Aspects of Convex Geometry Polyhedra, Linear Programming, Shellings, Voronoi Diagrams, Delaunay Triangulations

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

June 10, 2022

Aspects of Convex Geometry Polyhedra, Linear Programming, Shellings, Voronoi Diagrams, Delaunay Triangulations

Jean Gallier and Jocelyn Quaintance

Abstract: Some basic mathematical tools such as convex sets, polytopes and combinatorial topology, are used quite heavily in applied fields such as geometric modeling, meshing, computer vision, medical imaging and robotics. This report may be viewed as a tutorial and a set of notes on convex sets, polytopes, polyhedra, combinatorial topology, Voronoi Diagrams and Delaunay Triangulations. It is intended for a broad audience of mathematically inclined readers.

One of my (selfish!) motivations in writing these notes was to understand the concept of *shelling* and how it is used to prove the famous Euler-Poincaré formula (Poincaré, 1899) and the more recent *Upper Bound Theorem* (McMullen, 1970) for polytopes. Another of my motivations was to give a "correct" account of Delaunay triangulations and Voronoi diagrams in terms of (direct and inverse) stereographic projections onto a sphere and prove rigorously that the projective map that sends the (projective) sphere to the (projective) paraboloid works correctly, that is, maps the Delaunay triangulation and Voronoi diagram w.r.t. the lifting onto the sphere to the Delaunay diagram and Voronoi diagrams w.r.t. the traditional lifting onto the paraboloid. Here, the problem is that this map is only well defined (total) in projective space and we are forced to define the notion of convex polyhedron in projective space.

It turns out that in order to achieve (even partially) the above goals, I found that it was necessary to include quite a bit of background material on convex sets, polytopes, polyhedra and projective spaces. I have included a rather thorough treatment of the equivalence of $\mathcal{V}$-polytopes and $\mathcal{H}$-polytopes and also of the equivalence of $\mathcal{V}$-polyhedra and $\mathcal{H}$-polyhedra, which is a bit harder. In particular, the *Fourier-Motzkin elimination* method (a version of Gaussian elimination for inequalities) is discussed in some detail. I also had to include some material on projective spaces, projective maps and polar duality w.r.t. a nondegenerate quadric in order to define a suitable notion of "projective polyhedron" based on cones. To the best of our knowledge, this notion of projective polyhedron is new. We also believe that some of our proofs establishing the equivalence of $\mathcal{V}$-polyhedra and $\mathcal{H}$-polyhedra are new.

Since Chapters 2, 3, 4, and 5 contain all the background (and more) needed to discuss linear programming (including the simplex algorithm and duality), we have included some chapters on linear programming.

Key-words: Convex sets, polytopes, polyhedra, linear programming, simplex algorithm, strong duality, shellings, combinatorial topology, Voronoi diagrams, Delaunay triangulations.

Contents

Chapter 1

Introduction

1.1 Motivations and Goals

For the past eight years or so I have been teaching a graduate course whose main goal is to expose students to some fundamental concepts of geometry, keeping in mind their applications to geometric modeling, meshing, computer vision, medical imaging, robotics, *etc.* The audience has been primarily computer science students but a fair number of mathematics students and also students from other engineering disciplines (such as Electrical, Systems, Mechanical and Bioengineering) have been attending my classes. In the past three years, I have been focusing more on convexity, polytopes and combinatorial topology, as concepts and tools from these areas have been used increasingly in meshing and also in computational biology and medical imaging. One of my (selfish!) motivations was to understand the concept of *shelling* and how it is used to prove the famous Euler-Poincaré formula (Poincaré, 1899) and the more recent *Upper Bound Theorem* (McMullen, 1970) for polytopes. Another of my motivations was to give a "correct" account of Delaunay triangulations and Voronoi diagrams in terms of (direct and inverse) stereographic projections onto a sphere and prove rigorously that the projective map that sends the (projective) sphere to the (projective) paraboloid works correctly, that is, maps the Delaunay triangulation and Voronoi diagram w.r.t. the lifting onto the sphere to the Delaunay triangulation and Voronoi diagram w.r.t. the lifting onto the paraboloid. Moreover, the projections of these polyhedra onto the hyperplane $x_{d+1} = 0$, from the sphere or from the paraboloid, are identical. Here, the problem is that this map is only well defined (total) in projective space and we are forced to define the notion of convex polyhedron in projective space.

It turns out that in order to achieve (even partially) the above goals, I found that it was necessary to include quite a bit of background material on convex sets, polytopes, polyhedra and projective spaces. I have included a rather thorough treatment of the equivalence of $\mathcal{V}$-polytopes and $\mathcal{H}$-polytopes and also of the equivalence of $\mathcal{V}$-polyhedra and $\mathcal{H}$-polyhedra, which is a bit harder. In particular, the *Fourier-Motzkin elimination* method (a version of Gaussian elimination for inequalities) is discussed in some detail. I also had to include some material on projective spaces, projective maps and polar duality w.r.t. a nondegenerate

quadric, in order to define a suitable notion of "projective polyhedron" based on cones. This notion turned out to be indispensible to give a correct treatment of the Delaunay and Voronoi complexes using inverse stereographic projection onto a sphere and to prove rigorously that the well known projective map between the sphere and the paraboloid maps the Delaunay triangulation and the Voronoi diagram w.r.t. the sphere to the more traditional Delaunay triangulation and Voronoi diagram w.r.t. the paraboloid. To the best of our knowledge, this notion of projective polyhedron is new. We also believe that some of our proofs establishing the equivalence of $\mathcal{V}$-polyhedra and $\mathcal{H}$-polyhedra are new.

Chapter 10 on combinatorial topology is hardly original. However, most texts covering this material are either old fashion or too advanced. Yet, this material is used extensively in meshing and geometric modeling. We tried to give a rather intuitive yet rigorous exposition. We decided to introduce the terminology *combinatorial manifold*, a notion usually referred to as *triangulated manifold*.

A recurring theme in these notes is the process of "conification" (algebraically, "homogenization"), that is, forming a cone from some geometric object. Indeed, "conification" turns an object into a set of lines, and since lines play the role of points in projective geometry, "conification" ("homogenization") is the way to "projectivize" geometric affine objects. Then, these (affine) objects appear as "conic sections" of cones by hyperplanes, just the way the classical conics (ellipse, hyperbola, parabola) appear as conic sections.

It is worth warning our readers that convexity and polytope theory is deceptively simple. This is a subject where most intuitive propositions fail as soon as the dimension of the space is greater than 3 (definitely 4), because our human intuition is not very good in dimension greater than 3. Furthermore, rigorous proofs of seemingly very simple facts are often quite complicated and may require sophisticated tools (for example, shellings, for a correct proof of the Euler-Poincaré formula). Nevertheless, readers are urged to strenghten their geometric intuition; they should just be very vigilant! This is another case where Tate's famous saying is more than pertinent: "Reason geometrically, prove algebraically."

At first, these notes were meant as a complement to Chapter 3 (Properties of Convex Sets: A Glimpse) of my book (*Geometric Methods and Applications*, [30]). However, they turn out to cover much more material. For the reader's convenience, I have included Chapter 2 on affine geometry, and Chapter 3 (both from my book [30]) as part of Chapter 3 of these notes.

Since Chapters 2, 3, 4, and 5 contain all the background (and more) needed to discuss linear programming (including the simplex algorithm and duality), we have included some chapters on linear programming.

Most of the material on convex sets is taken from Berger [8] (*Geometry II*). Other relevant sources include Ziegler [69], Grünbaum [36] Barvinok [4], Valentine [65], Rockafellar [51], Bourbaki (Topological Vector Spaces) [13], and Lax [40], the last four dealing with affine spaces of infinite dimension. As to polytopes and polyhedra, "the" classic reference is

Grünbaum [36]. Other good references include Ziegler [69], Ewald [26], Cromwell [22], and Thomas [62].

The recent book by Thomas contains an excellent and easy going presentation of polytope theory. This book also gives an introduction to the theory of triangulations of point configurations, including the definition of secondary polytopes and state polytopes, which happen to play a role in certain areas of biology. For this, a quick but very efficient presentation of Gröbner bases is provided. We highly recommend Thomas's book [62] as further reading. It is also an excellent preparation for the more advanced book by Sturmfels [61]. However, in our opinion, the "bible" on polytope theory is without any contest, Ziegler [69], a masterly and beautiful piece of mathematics. In fact, our Chapter 11 is heavily inspired by Chapter 8 of Ziegler. However, the pace of Ziegler's book is quite brisk and we hope that our more pedestrian account will inspire readers to go back and read the masters.

In a not too distant future, I would like to write about constrained Delaunay triangulations, a formidable topic, please be patient!

I wish to thank Marcelo Siqueira for catching many typos and mistakes and for his many helpful suggestions regarding the presentation. At least a third of this manuscript was written while I was on sabbatical at INRIA, Sophia Antipolis, in the Asclepios Project. My deepest thanks to Nicholas Ayache and his colleagues (especially Xavier Pennec and Hervé Delingette) for inviting me to spend a wonderful and very productive year and for making me feel perfectly at home within the Asclepios Project.

Chapter 2

Basics of Affine Geometry

L'algèbre n'est qu'une géométrie écrite; la géométrie n'est qu'une algèbre figurée.
—**Sophie Germain**

2.1 Affine Spaces

Geometrically, curves and surfaces are usually considered to be sets of points with some special properties, living in a space consisting of "points." Typically, one is also interested in geometric properties invariant under certain transformations, for example, translations, rotations, projections, etc. One could model the space of points as a vector space, but this is not very satisfactory for a number of reasons. One reason is that the point corresponding to the zero vector (0), called the origin, plays a special role, when there is really no reason to have a privileged origin. Another reason is that certain notions, such as parallelism, are handled in an awkward manner. But the deeper reason is that vector spaces and affine spaces really have different geometries. The geometric properties of a vector space are invariant under the group of bijective linear maps, whereas the geometric properties of an affine space are invariant under the group of bijective affine maps, and these two groups are not isomorphic. Roughly speaking, there are more affine maps than linear maps.

Affine spaces provide a better framework for doing geometry. In particular, it is possible to deal with points, curves, surfaces, etc., in an **intrinsic manner**, that is, independently of any specific choice of a coordinate system. As in physics, this is highly desirable to really understand what is going on. Of course, coordinate systems have to be chosen to finally carry out computations, but one should learn to resist the temptation to resort to coordinate systems until it is really necessary.

Affine spaces are the right framework for dealing with motions, trajectories, and physical forces, among other things. Thus, affine geometry is crucial to a clean presentation of kinematics, dynamics, and other parts of physics (for example, elasticity). After all, a rigid motion is an affine map, but not a linear map in general. Also, given an $m \times n$ matrix A

and a vector $b \in \mathbb{R}^m$, the set $U = \{x \in \mathbb{R}^n \mid Ax = b\}$ of solutions of the system $Ax = b$ is an affine space, but not a vector space (linear space) in general.

Use coordinate systems only when needed!

This chapter proceeds as follows. We take advantage of the fact that almost every affine concept is the counterpart of some concept in linear algebra. We begin by defining affine spaces, stressing the physical interpretation of the definition in terms of points (particles) and vectors (forces). Corresponding to linear combinations of vectors, we define affine combinations of points (barycenters), realizing that we are forced to restrict our attention to families of scalars adding up to 1. Corresponding to linear subspaces, we introduce affine subspaces as subsets closed under affine combinations. Then, we characterize affine subspaces in terms of certain vector spaces called their directions. This allows us to define a clean notion of parallelism. Next, corresponding to linear independence and bases, we define affine independence and affine frames. We also define convexity. Corresponding to linear maps, we define affine maps as maps preserving affine combinations. We show that every affine map is completely defined by the image of one point and a linear map. Then, we investigate briefly some simple affine maps, the translations and the central dilatations. At this point, we give a glimpse of affine geometry. We prove the theorems of Thales, Pappus, and Desargues. After this, the definition of affine hyperplanes in terms of affine forms is reviewed. The section ends with a closer look at the intersection of affine subspaces.

Our presentation of affine geometry is far from being comprehensive, and it is biased toward the algorithmic geometry of curves and surfaces. For more details, the reader is referred to Pedoe [48], Snapper and Troyer [55], Berger [7, 8], Coxeter [21], Samuel [52], Tisseron [64], Fresnel [27], Vienne [67], and Hilbert and Cohn-Vossen [37].

Suppose we have a particle moving in 3D space and that we want to describe the trajectory of this particle. If one looks up a good textbook on dynamics, such as Greenwood [35], 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. Curiously, the notion of a frame is rarely defined precisely, but it is easy to infer that 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 by $\overrightarrow{ab}$, from a to b, the vector $\overrightarrow{ab} = (b_1 - a_1, b_2 - a_2, b_3 - a_3)$. Note that

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

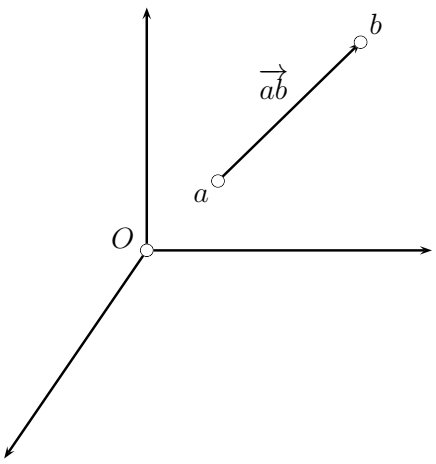


Figure 2.1: Points and free vectors.

addition being understood as addition in $\mathbb{R}^3$. Then, in the standard frame, given a point $x = (x_1, x_2, x_3)$, the position of x is the vector $\overrightarrow{Ox} = (x_1, x_2, x_3)$, which coincides with the point itself. In the standard frame, points and vectors are identified. Points and free vectors are illustrated in Figure 2.1.

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:

$$\overrightarrow{Ox} = (x_1, x_2, x_3)$$

in the frame $(O, (e_1, e_2, e_3))$ and

$$\overrightarrow{\Omega x} = (x_1 - \omega_1, x_2 - \omega_2, x_3 - \omega_3)$$

in the frame $(\Omega, (e_1, e_2, e_3))$. See Figure 2.2.

This is because

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

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. It may be computationally convenient to deal with points using position vectors, but such a treatment is not frame invariant, which has undesirable effects.

Inspired by physics, we deem it 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. First, let us review the notion of linear combination of vectors. Given two vectors u and v of coordinates (u_1, u_2, u_3) and (v_1, v_2, v_3) with respect

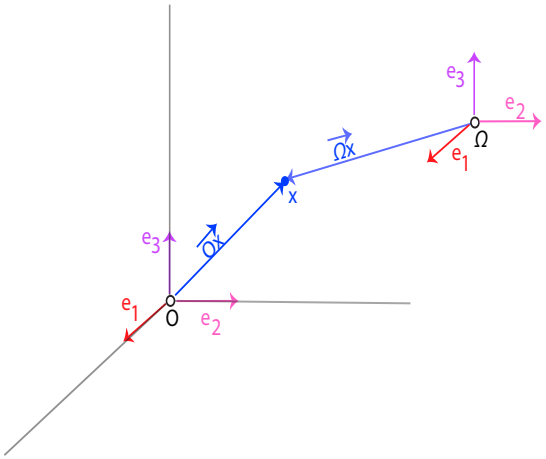


Figure 2.2: The two position vectors for the point x.

to the basis (e_1, e_2, e_3), for any two scalars λ, μ, we can define the linear combination $\lambda u + \mu v$ as the vector of coordinates

$$(\lambda u_1 + \mu v_1, \lambda u_2 + \mu v_2, \lambda u_3 + \mu v_3).$$

If we choose a different basis (e'_1, e'_2, e'_3) and if the matrix P expressing the vectors (e'_1, e'_2, e'_3) over the basis (e_1, e_2, e_3) is

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

which means that the columns of P are the coordinates of the e'_j over the basis (e_1, e_2, e_3), since

$$u_1 e_1 + u_2 e_2 + u_3 e_3 = u'_1 e'_1 + u'_2 e'_2 + u'_3 e'_3$$

and

$$v_1 e_1 + v_2 e_2 + v_3 e_3 = v'_1 e'_1 + v'_2 e'_2 + v'_3 e'_3,$$

it is easy to see that the coordinates (u_1, u_2, u_3) and (v_1, v_2, v_3) of u and v with respect to the basis (e_1, e_2, e_3) are given in terms of the coordinates (u'_1, u'_2, u'_3) and (v'_1, v'_2, v'_3) of u and v with respect to the basis (e'_1, e'_2, e'_3) by the matrix equations

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = P \begin{pmatrix} u'_1 \\ u'_2 \\ u'_3 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = P \begin{pmatrix} v'_1 \\ v'_2 \\ v'_3 \end{pmatrix}.$$

From the above, we get

$$\begin{pmatrix} u'_1 \\ u'_2 \\ u'_3 \end{pmatrix} = P^{-1} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} v'_1 \\ v'_2 \\ v'_3 \end{pmatrix} = P^{-1} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

and by linearity, the coordinates

$$(\lambda u'_1 + \mu v'_1, \lambda u'_2 + \mu v'_2, \lambda u'_3 + \mu v'_3)$$

of $\lambda u + \mu v$ with respect to the basis (e'_1, e'_2, e'_3) are given by

$$\begin{pmatrix} \lambda u'_1 + \mu v'_1 \\ \lambda u'_2 + \mu v'_2 \\ \lambda u'_3 + \mu v'_3 \end{pmatrix} = \lambda P^{-1} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} + \mu P^{-1} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = P^{-1} \begin{pmatrix} \lambda u_1 + \mu v_1 \\ \lambda u_2 + \mu v_2 \\ \lambda u_3 + \mu v_3 \end{pmatrix}.$$

Everything worked out because the change of basis does not involve a change of origin. On the other hand, 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 $\overrightarrow{O\Omega} = (\omega_1, \omega_2, \omega_3)$, given two points a, 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) 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

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

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 $\lambda + \mu = 1$. See Figure 2.3.

Thus, we have 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.

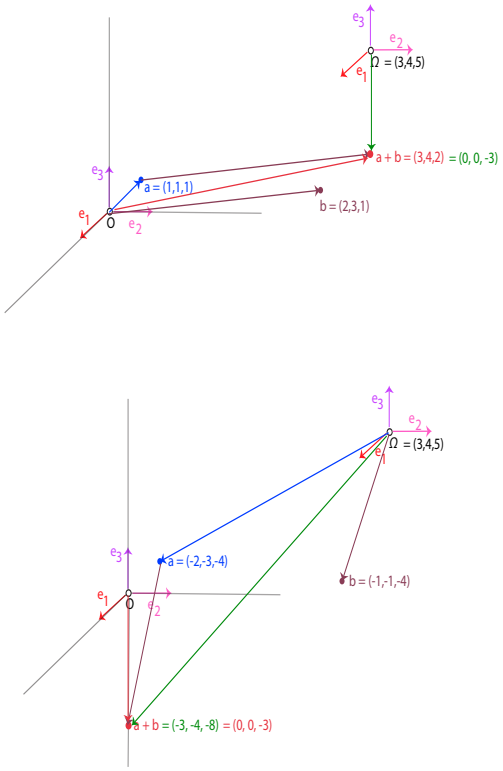


Figure 2.3: The top figure shows the location of the "point" sum $a + b$ with respect to the frame $(O, (e_1, e_2, e_3))$, while the bottom figure shows the location of the "point" sum $a + b$ with respect to the frame $(\Omega, (e_1, e_2, e_3))$.

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. We can imagine that v is placed such that its origin coincides with a and that its tip coincides with b. This action $+\colon \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ satisfies some crucial properties. For example,

$$\begin{aligned} a + 0 &= a, \\ (a + u) + v &= a + (u + v), \end{aligned}$$

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

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

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). The basic idea is to consider two (distinct) sets E and $\overrightarrow{E}$, where E is a set of points (with no structure) and $\overrightarrow{E}$ is a vector space (of free vectors) acting on the set E.

Did you say "A fine space"?

Intuitively, we can think of the elements of $\overrightarrow{E}$ as forces moving the points in E, considered as physical particles. The effect of applying a force (free vector) $u \in \overrightarrow{E}$ to a point $a \in E$ is a translation. By this, we mean that for every force $u \in \overrightarrow{E}$, the action of the force u is to "move" every point $a \in E$ to the point $a + u \in E$ obtained by the translation corresponding to u viewed as a vector. Since translations can be composed, it is natural that $\overrightarrow{E}$ is a vector space.

For simplicity, it is assumed that all vector spaces under consideration are defined over the field $\mathbb{R}$ of real numbers. Most of the definitions and results also hold for an arbitrary field K, although some care is needed when dealing with fields of characteristic different from zero. 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 of scalars are finite. The formal definition of an affine space is as follows.

Definition 2.1. An *affine space* is either the degenerate space reduced to 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 $+ \colon E \times \overrightarrow{E} \to E$, 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 by $\overrightarrow{ab}$, or sometimes by $\mathbf{ab}$, or even by $b - a$. Thus, we also write

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

(or $b = a + \mathbf{ab}$, 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 $\dim(E)$.

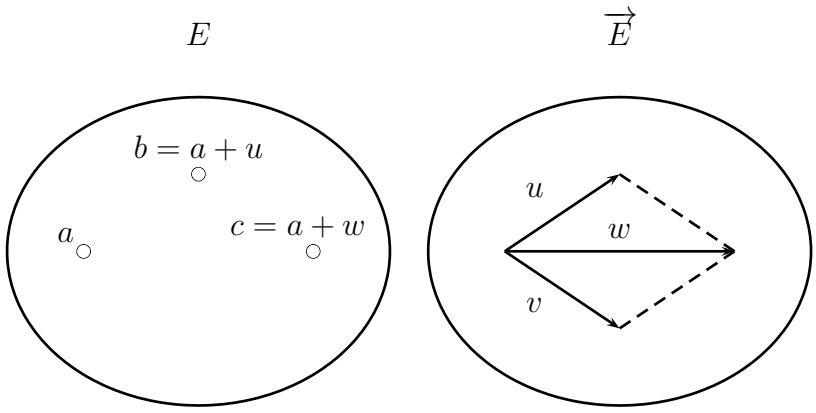


Figure 2.4: Intuitive picture of an affine space.

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

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

for all $a \in E$ and all $v \in \overrightarrow{E}$, since $\overrightarrow{a(a+v)}$ is the unique vector such that $a+v = a+\overrightarrow{a(a+v)}$. Thus, $b = a + v$ is equivalent to $\overrightarrow{ab} = v$. Figure 2.4 gives an intuitive picture of an affine space. It is natural to think of all vectors as having the same origin, the null vector.

The axioms defining an affine space $\langle E, \overrightarrow{E}, + \rangle$ can be interpreted intuitively as saying that E and $\overrightarrow{E}$ are two different ways of looking at the same object, but wearing different sets of glasses, the second set of glasses depending on the choice of an "origin" in E. Indeed, we can choose to look at the points in E, forgetting that every pair (a, b) of points defines a unique vector $\overrightarrow{ab}$ in $\overrightarrow{E}$, or we can choose to look at the vectors u in $\overrightarrow{E}$, forgetting the points in E. Furthermore, if we also pick any point a in E, a point that can be viewed as an *origin* in E, then we can recover all the points in E as the translated points $a + u$ for all $u \in \overrightarrow{E}$. This can be formalized by defining two maps between E and $\overrightarrow{E}$.

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

$$u \mapsto a + u,$$

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

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

where $b \in E$. The composition of the first mapping with the second is

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

which, in view of (A3), yields u. The composition of the second with the first mapping is

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

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 with $\overrightarrow{E}$ via the mapping $b \mapsto \overrightarrow{ab}$, we say that we consider E as the vector space obtained *by taking a as the origin in E*, and we denote it by E_a. Because E_a is a vector space, to be consistent with our notational conventions we should use the notation $\overrightarrow{E_a}$ (using an arrow), instead of E_a. However, for simplicity, we stick to the notation 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. Nevertheless, as soon as we commit to an origin a in E, we can view E as the vector space E_a. However, we urge the reader to think of E as a physical set of points and of $\overrightarrow{E}$ as a set of forces acting on E, rather than reducing E to some isomorphic copy of $\mathbb{R}^n$. After all, points are points, and not vectors! For notational simplicity, we will often denote an affine space $\langle E, \overrightarrow{E}, + \rangle$ by $(E, \overrightarrow{E})$, or even by 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 subtracted (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 $\overrightarrow{E}$ 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 by $\mathbb{A}^n$. In general, if K is any field, the affine space $\langle K^n, K^n, + \rangle$ is denoted by $\mathbb{A}^n_K$. In order to distinguish between the double role played by members of $\mathbb{R}^n$, points and vectors, we will denote points by row vectors, and vectors by 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 by $\mathbf{x}$ (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 $n = 1, 2, 3$.

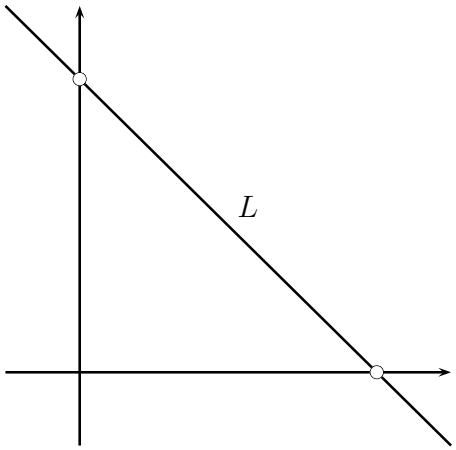


Figure 2.5: An affine space: the line of equation $x + y - 1 = 0$

2.2 Examples of Affine Spaces

Let us now give an example of an affine space that is not given as a vector space (at least, not in an obvious fashion). Consider the subset L of $\mathbb{A}^2$ consisting of all points (x, y) satisfying the equation

$$x + y - 1 = 0.$$

The set L is the line of slope -1 passing through the points $(1, 0)$ and $(0, 1)$ shown in Figure 2.5.

The line L can be made into an official affine space by defining the action $+\colon L \times \mathbb{R} \to L$ of $\mathbb{R}$ on L defined such that for every point $(x, 1 - x)$ on L and any $u \in \mathbb{R}$,

$$(x, 1 - x) + u = (x + u, 1 - x - u).$$

It is immediately verified that this action makes L into an affine space. For example, for any two points $a = (a_1, 1 - a_1)$ and $b = (b_1, 1 - b_1)$ on L, the unique (vector) $u \in \mathbb{R}$ such that $b = a + u$ is $u = b_1 - a_1$. Note that the vector space $\mathbb{R}$ is isomorphic to the line of equation $x + y = 0$ passing through the origin.

Similarly, consider the subset H of $\mathbb{A}^3$ consisting of all points (x, y, z) satisfying the equation

$$x + y + z - 1 = 0.$$

The set H is the plane passing through the points $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. The plane H can be made into an official affine space by defining the action $+\colon H \times \mathbb{R}^2 \to H$ of $\mathbb{R}^2$ on

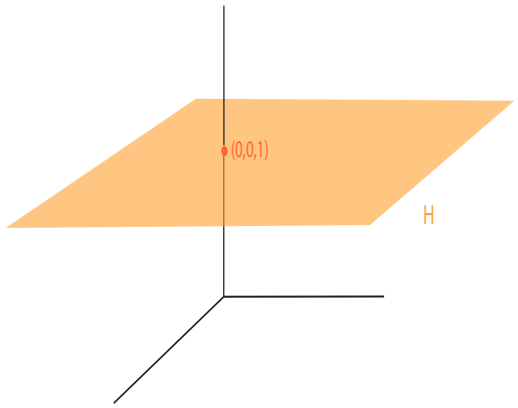


Figure 2.6: An affine space: the plane $x + y + z - 1 = 0$.

H defined such that for every point $(x, y, 1 - x - y)$ on H and any $\begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^2$,

$$(x, y, 1 - x - y) + \begin{pmatrix} u \\ v \end{pmatrix} = (x + u, y + v, 1 - x - u - y - v).$$

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 $\begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^2$,

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

This should dispel any idea that affine spaces are dull. Affine spaces not already equipped with an obvious vector space structure arise in projective geometry.

2.3 Chasles's Identity

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

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

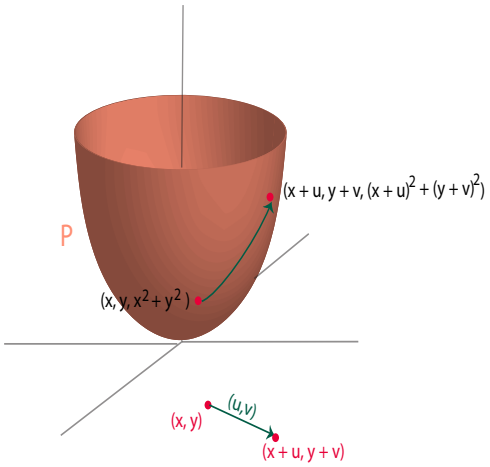


Figure 2.7: The paraboloid of revolution P viewed as a two-dimensional affine space.

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

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

which is known as *Chasles's identity*, and illustrated in Figure 2.8.

Since $a = a + \overrightarrow{aa}$ and by (A1) $a = a + 0$, by (A3) we get

$$\overrightarrow{aa} = 0.$$

Thus, letting $a = c$ in Chasles's identity, we get

$$\overrightarrow{ba} = -\overrightarrow{ab}.$$

Given any four points $a, b, c, d \in E$, since by Chasles's identity

$$\overrightarrow{ab} + \overrightarrow{bc} = \overrightarrow{ad} + \overrightarrow{dc} = \overrightarrow{ac},$$

we have the *parallelogram law*

$$\overrightarrow{ab} = \overrightarrow{dc} \quad \text{iff} \quad \overrightarrow{bc} = \overrightarrow{ad}.$$

2.4 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, as we saw in Section 2.1. Since this problem is the reason for introducing affine combinations, at the risk

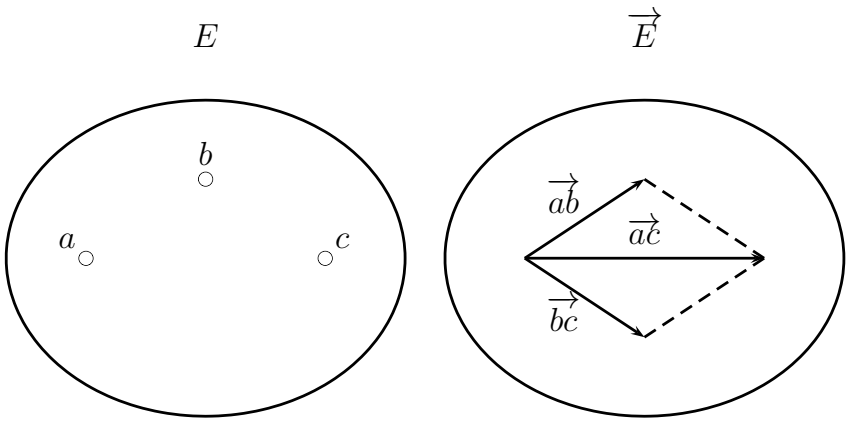


Figure 2.8: Points and corresponding vectors in affine geometry.

of boring certain readers, we give another example showing what goes wrong if we are not careful in defining linear combinations of points.

Consider $\mathbb{R}^2$ as an affine space, under its natural coordinate system with origin $O = (0, 0)$ and basis vectors $\begin{pmatrix}1\\0\end{pmatrix}$ and $\begin{pmatrix}0\\1\end{pmatrix}$. Given any two points $a = (a_1, a_2)$ and $b = (b_1, b_2)$, it is natural to define the affine combination $\lambda a + \mu b$ as the point of coordinates

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

Thus, when $a = (-1, -1)$ and $b = (2, 2)$, the point $a + b$ is the point $c = (1, 1)$.

Let us now consider the new coordinate system with respect to the origin $c = (1, 1)$ (and the same basis vectors). This time, the coordinates of a are $(-2, -2)$, the coordinates of b are $(1, 1)$, and the point $a+b$ is the point d of coordinates $(-1, -1)$. However, it is clear that the point d is identical to the origin $O = (0, 0)$ of the first coordinate system. This situation is illustrated in Figure 2.9.

Thus, $a + b$ corresponds to two different points depending on which coordinate system is used for its computation!

This shows that 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 proposition shows.

Proposition 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

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

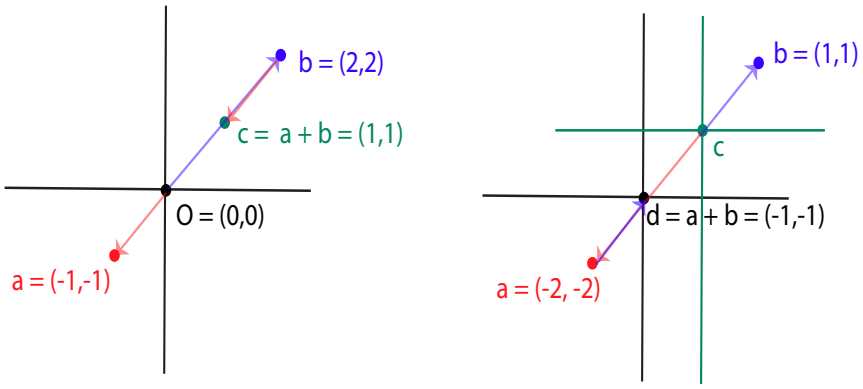


Figure 2.9: The example from the beginning of Section 2.4.

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

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

Proof. (1) By Chasles's identity (see Section 2.3), we have

$$\begin{aligned} a + \sum_{i\in I}\lambda_i\overrightarrow{aa_i} &= a + \sum_{i\in I}\lambda_i(\overrightarrow{ab} + \overrightarrow{ba_i}) \\ &= a + \left(\sum_{i\in I}\lambda_i\right)\overrightarrow{ab} + \sum_{i\in I}\lambda_i\overrightarrow{ba_i} \\ &= a + \overrightarrow{ab} + \sum_{i\in I}\lambda_i\overrightarrow{ba_i} && \text{since } \textstyle\sum_{i\in I}\lambda_i = 1 \\ &= b + \sum_{i\in I}\lambda_i\overrightarrow{ba_i} && \text{since } b = a + \overrightarrow{ab}. \end{aligned}$$

An illustration of this calculation in $\mathbb{A}^2$ is provided by Figure 2.10.

(2) We also have

$$\begin{aligned} \sum_{i\in I}\lambda_i\overrightarrow{aa_i} &= \sum_{i\in I}\lambda_i(\overrightarrow{ab} + \overrightarrow{ba_i}) \\ &= \left(\sum_{i\in I}\lambda_i\right)\overrightarrow{ab} + \sum_{i\in I}\lambda_i\overrightarrow{ba_i} \\ &= \sum_{i\in I}\lambda_i\overrightarrow{ba_i}, \end{aligned}$$

since $\sum_{i\in I}\lambda_i = 0$. □

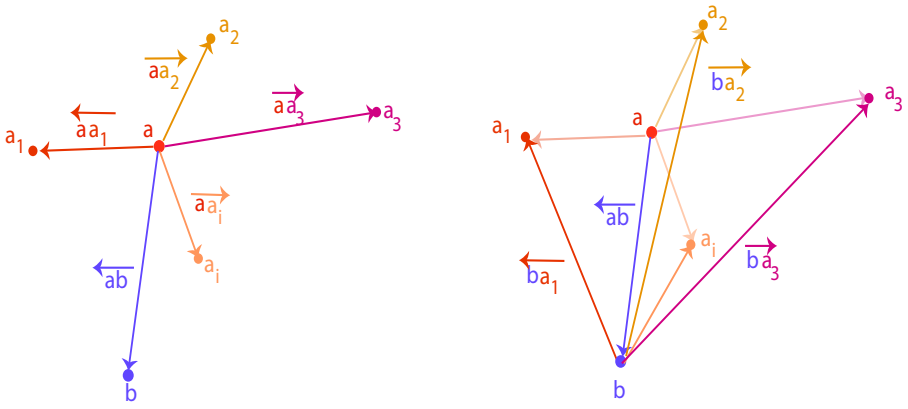


Figure 2.10: Part (1) of Proposition 2.1.

Thus, by Proposition 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

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

is independent of the choice of the origin $a \in E$. This property motivates the following definition.

Definition 2.2. For any family of points $(a_i)_{i\in I}$ in E, for any family $(\lambda_i)_{i\in I}$ of scalars such that $\sum_{i\in I}\lambda_i = 1$, and for any $a \in E$, the point

$$a + \sum_{i\in I}\lambda_i\overrightarrow{aa_i}$$

(which is independent of $a \in E$, by Proposition 2.1) is called the *barycenter (or barycentric combination, or affine combination) of the points a_i assigned the weights λ_i*, and it is denoted by

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

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 the unique point such that

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

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

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

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).

Remarks:

(1) Since the barycenter of a family $((a_i, \lambda_i))_{i \in I}$ of weighted points is defined for families $(\lambda_i)_{i \in I}$ of scalars with finite support (and such that $\sum_{i \in I} \lambda_i = 1$), we might as well assume that I is finite. Then, for all $m \geq 2$, it is easy to prove that the barycenter of m weighted points can be obtained by repeated computations of barycenters of two weighted points.

(2) This result still holds, provided that the field K has at least three distinct elements, but the proof is trickier!

(3) When $\sum_{i \in I} \lambda_i = 0$, the vector $\sum_{i \in I} \lambda_i \overrightarrow{aa_i}$ does not depend on the point a, and we may denote it by $\sum_{i \in I} \lambda_i a_i$. This observation will be used to define a vector space in which linear combinations of both points and vectors make sense, regardless of the value of $\sum_{i \in I} \lambda_i$.

Figure 2.11 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 $(c, 1)$.

The point g_1 can be constructed geometrically as the middle of the segment joining c to the middle $\frac{1}{2}a + \frac{1}{2}b$ of the segment (a, b), since

$$g_1 = \frac{1}{2}\left(\frac{1}{2}a + \frac{1}{2}b\right) + \frac{1}{2}c.$$

The point g_2 can be constructed geometrically as the point such that the middle $\frac{1}{2}b + \frac{1}{2}c$ of the segment (b, c) is the middle of the segment (a, g_2), since

$$g_2 = -a + 2\left(\frac{1}{2}b + \frac{1}{2}c\right).$$

Later on, we will see that a polynomial curve can be defined as a set of barycenters of a fixed number of points. For example, let (a, b, c, d) be a sequence of points in $\mathbb{A}^2$. Observe that

$$(1-t)^3 + 3t(1-t)^2 + 3t^2(1-t) + t^3 = 1,$$

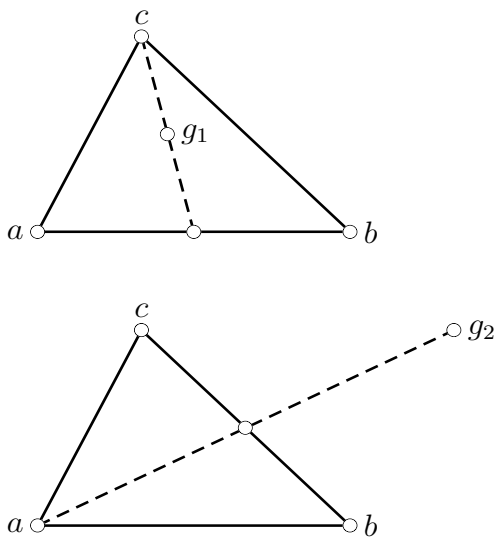


Figure 2.11: Barycenters, $g_1 = \frac{1}{4}a + \frac{1}{4}b + \frac{1}{2}c, \quad g_2 = -a + b + c$

since the sum on the left-hand side is obtained by expanding $(t + (1-t))^3 = 1$ using the binomial formula. Thus,

$$(1-t)^3\, a + 3t(1-t)^2\, b + 3t^2(1-t)\, c + t^3\, d$$

is a well-defined affine combination. Then, we can define the curve $F\colon \mathbb{A} \to \mathbb{A}^2$ such that

$$F(t) = (1-t)^3\, a + 3t(1-t)^2\, b + 3t^2(1-t)\, c + t^3\, d.$$

Such a curve is called a *Bézier curve*, and (a, b, c, d) are called its *control points*. Note that the curve passes through a and d, but generally not through b and c. It can be sbown that any point $F(t)$ on the curve can be constructed using an algorithm performing affine interpolation steps (the *de Casteljau algorithm*).

2.5 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. 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 weighted points $((a_i, \lambda_i))_{i \in I}$ in V 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, 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

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

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

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

Indeed, $(x_i, y_i) \in U$ means that

$$ax_i + by_i = c,$$

and if we multiply both sides of this equation by λ_i and add up the resulting m equations, we get

$$\sum_{i=1}^{m} (\lambda_i a x_i + \lambda_i b y_i) = \sum_{i=1}^{m} \lambda_i c,$$

and since $\lambda_1 + \cdots + \lambda_m = 1$, we get

$$a\left(\sum_{i=1}^{m} \lambda_i x_i\right) + b\left(\sum_{i=1}^{m} \lambda_i y_i\right) = \left(\sum_{i=1}^{m} \lambda_i\right) c = c,$$

which shows that

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

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

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

i.e., the set of solutions of the homogeneous equation

$$ax + by = 0$$

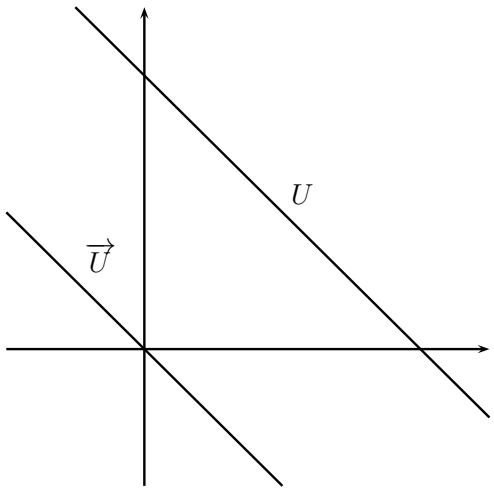


Figure 2.12: An affine line U and its direction.

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 $\mathbb{A}^2$, a line that is parallel to the line U of equation $ax + by = c$, as illustrated in Figure 2.12.

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

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

where

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

First, $(x_0, y_0) + \overrightarrow{U} \subseteq U$, since $ax_0 + by_0 = c$ and $au_1 + bu_2 = 0$ for all $(u_1, u_2) \in \overrightarrow{U}$. Second, if $(x, y) \in U$, then $ax + by = c$, and since we also have $ax_0 + by_0 = c$, by subtraction, we get

$$a(x - x_0) + b(y - y_0) = 0,$$

which shows that $(x - x_0, y - y_0) \in \overrightarrow{U}$, and thus $(x, y) \in (x_0, y_0) + \overrightarrow{U}$. Hence, we also have $U \subseteq (x_0, y_0) + \overrightarrow{U}$, and $U = (x_0, y_0) + \overrightarrow{U}$.

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 $(x_0, y_0) \in U$,

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

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 $A\mathbf{x} = b$. For the sake of minimizing the amount of notation, we stick to the simpler (yet incorrect) notation $Ax = b$. 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 be characterized in terms of subspaces of $\overrightarrow{E}$. Let V be a nonempty subset of E. For every family $(a_1, \ldots, a_n)$ in V, for any family $(\lambda_1, \ldots, \lambda_n)$ of scalars, and for every point $a \in V$, observe that for every $x \in E$,

$$x = a + \sum_{i=1}^{n} \lambda_i \overrightarrow{aa_i}$$

is the barycenter of the family of weighted points

$$\left((a_1, \lambda_1), \ldots, (a_n, \lambda_n), \left(a, 1 - \sum_{i=1}^{n} \lambda_i\right)\right),$$

since

$$\sum_{i=1}^{n} \lambda_i + \left(1 - \sum_{i=1}^{n} \lambda_i\right) = 1.$$

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

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

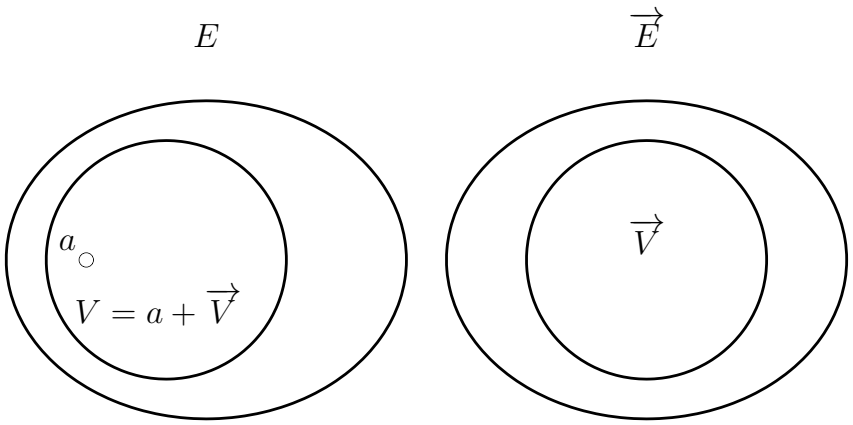


Figure 2.13: An affine subspace V and its direction $\overrightarrow{V}$.

Proposition 2.2. *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

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

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

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

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 $\overrightarrow{E}$ and for any $a \in E$, the set $V = a + \overrightarrow{V}$ is an affine subspace.

Proof. The proof is straightforward, and is omitted. It is also given in Gallier [29]. □

In particular, when E is the natural affine space associated with a vector space $\overrightarrow{E}$, Proposition 2.2 shows that every affine subspace of E is of the form $u + \overrightarrow{U}$, for a subspace $\overrightarrow{U}$ of $\overrightarrow{E}$. The subspaces of $\overrightarrow{E}$ are the affine subspaces of E that contain 0.

The subspace $\overrightarrow{V}$ associated with an affine subspace V is called the *direction of V*. It is also clear that the map $+ \colon V \times \overrightarrow{V} \to V$ induced by $+ \colon E \times \overrightarrow{E} \to E$ confers to $\langle V, \overrightarrow{V}, + \rangle$ an affine structure. Figure 2.13 illustrates the notion of affine subspace.

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 a *hyperplane* (recall that a subspace F of a vector space E has codimension 1 iff there is some subspace G of dimension 1 such that $E = F \oplus G$, the direct sum of F and G, see Strang [60] or Lang [38]).

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 $\overrightarrow{ab}$.

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).

By Proposition 2.2, a line is specified by a point $a \in E$ and a nonzero vector $v \in \overrightarrow{E}$, i.e., a line is the set of all points of the form $a + \lambda v$, for $\lambda \in \mathbb{R}$.

We say that three points a, b, c are *collinear* if the vectors $\overrightarrow{ab}$ and $\overrightarrow{ac}$ 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 $\overrightarrow{ac} = \lambda \overrightarrow{ab}$, and we define the ratio $\frac{\overrightarrow{ac}}{\overrightarrow{ab}} = \lambda$.

A plane is specified by a point $a \in E$ and two linearly independent vectors $u, v \in \overrightarrow{E}$, i.e., a plane is the set of all points of the form $a + \lambda u + \mu v$, for $\lambda, \mu \in \mathbb{R}$.

We say that four points a, b, c, d are *coplanar* if the vectors $\overrightarrow{ab}, \overrightarrow{ac}$, and $\overrightarrow{ad}$ are linearly dependent. Hyperplanes will be characterized a little later.

Proposition 2.3. *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 $(a_i)_{i \in I}$.*

Proof. If $(a_i)_{i \in I}$ is empty, then $V = \emptyset$, because of the condition $\sum_{i \in I} \lambda_i = 1$. If $(a_i)_{i \in I}$ is nonempty, then the smallest affine subspace containing $(a_i)_{i \in I}$ must contain the set V of barycenters $\sum_{i \in I} \lambda_i a_i$, and thus, it is enough to show that V is closed under affine combinations, which is immediately verified. □

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

Remarks:

(1) 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, then 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) This result still holds if the field K has at least three distinct elements, but the proof is trickier!

2.6 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 $(\overrightarrow{a_i a_j})_{j\in(I-\{i\})}$ of vectors obtained by choosing any a_i as an origin. First, the following proposition shows that it is sufficient to consider only one of these families.

Proposition 2.4. *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 $(\overrightarrow{a_i a_j})_{j\in(I-\{i\})}$ is linearly independent for some $i \in I$, then $(\overrightarrow{a_i a_j})_{j\in(I-\{i\})}$ is linearly independent for every $i \in I$.*

Proof. Assume that the family $(\overrightarrow{a_i a_j})_{j\in(I-\{i\})}$ is linearly independent for some specific $i \in I$. Let $k \in I$ with $k \neq i$, and assume that there are some scalars $(\lambda_j)_{j\in(I-\{k\})}$ such that

$$\sum_{j\in(I-\{k\})} \lambda_j \overrightarrow{a_k a_j} = 0.$$

Since

$$\overrightarrow{a_k a_j} = \overrightarrow{a_k a_i} + \overrightarrow{a_i a_j},$$

we have

$$\begin{aligned}
\sum_{j\in(I-\{k\})} \lambda_j \overrightarrow{a_k a_j} &= \sum_{j\in(I-\{k\})} \lambda_j \overrightarrow{a_k a_i} + \sum_{j\in(I-\{k\})} \lambda_j \overrightarrow{a_i a_j}, \\
&= \sum_{j\in(I-\{k\})} \lambda_j \overrightarrow{a_k a_i} + \sum_{j\in(I-\{i,k\})} \lambda_j \overrightarrow{a_i a_j}, \\
&= \sum_{j\in(I-\{i,k\})} \lambda_j \overrightarrow{a_i a_j} - \Big(\sum_{j\in(I-\{k\})} \lambda_j\Big) \overrightarrow{a_i a_k},
\end{aligned}$$

and thus

$$\sum_{j\in(I-\{i,k\})} \lambda_j \overrightarrow{a_i a_j} - \Big(\sum_{j\in(I-\{k\})} \lambda_j\Big) \overrightarrow{a_i a_k} = 0.$$

Since the family $(\overrightarrow{a_i a_j})_{j\in(I-\{i\})}$ is linearly independent, we must have $\lambda_j = 0$ for all $j \in (I - \{i, k\})$ and $\sum_{j\in(I-\{k\})} \lambda_j = 0$, which implies that $\lambda_j = 0$ for all $j \in (I - \{k\})$. □

We define affine independence as follows.

Definition 2.4. 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 $(\overrightarrow{a_i a_j})_{j\in(I-\{i\})}$ is linearly independent for some $i \in I$.

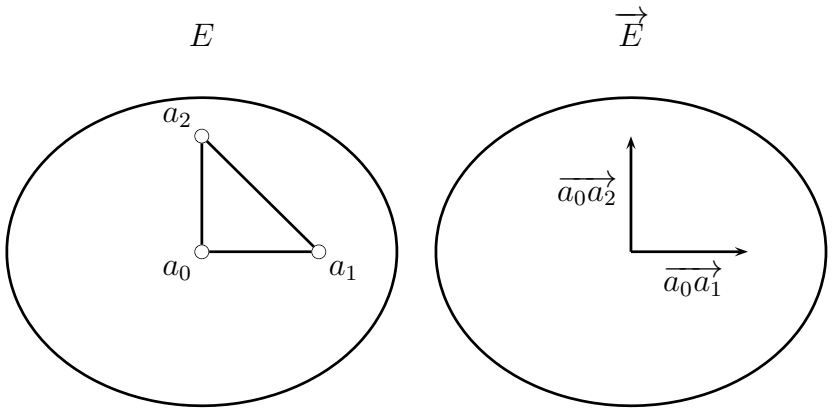


Figure 2.14: Affine independence and linear independence.

Definition 2.4 is reasonable, because by Proposition 2.4, the independence of the family $(\overrightarrow{a_ia_j})_{j\in(I-\{i\})}$ 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.

Proposition 2.5. *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 $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ is linearly independent.*

Proof. The proof is straightforward and is omitted. It is also given in Gallier [29]. □

Proposition 2.5 suggests the notion of affine frame. Affine frames are the affine analogues of bases in vector spaces. 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 $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ 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 $\overrightarrow{E}$. Figure 2.14 illustrates the notion of affine independence.

Remark: The above observation also applies to infinite families $(a_i)_{i\in I}$ of points in E and families $(\overrightarrow{u_i})_{i\in I-\{0\}}$ of vectors in $\overrightarrow{E}$, provided that the index set I contains 0.

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

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

The scalars $(x_1, \ldots, x_m)$ may be considered as coordinates with respect to $(a_0, (\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}))$. Since

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

$x \in E$ 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 $(a_0, \ldots, a_m)$. All this is summarized in the following definition.

Definition 2.5. 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 the list of vectors $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ is a basis of $\overrightarrow{E}$. The pair $(a_0, (\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}))$ is also called an *affine frame with origin* a_0. Then, every $x \in E$ can be expressed as

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

for a unique family $(x_1, \ldots, x_m)$ of scalars, called the *coordinates of x w.r.t. the affine frame* $(a_0, (\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}))$. 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* $(a_0, \ldots, a_m)$. See Figure 2.15.

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. A family $(a_i)_{i\in I}$ of points in E is *affinely dependent* if it is not affinely independent. We can also characterize affinely dependent families as follows.

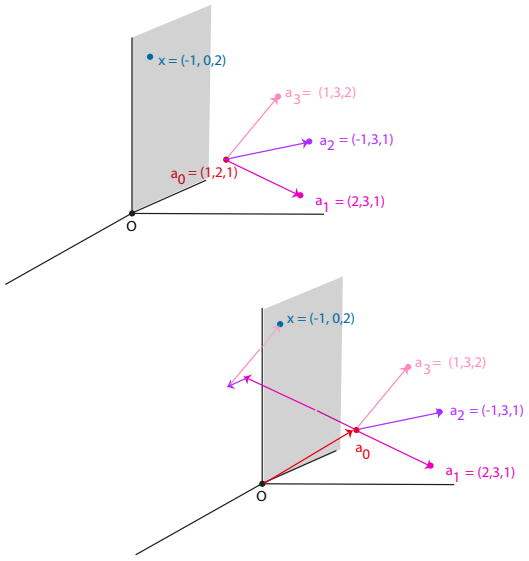


Figure 2.15: The affine frame (a_0, a_1, a_2, a_3) for $\mathbb{A}^3$. The coordinates for $x = (-1, 0, 2)$ are $x_1 = -8/3$, $x_2 = -1/3$, $x_3 = 1$, while the barycentric coordinates for x are $\lambda_0 = 3$, $\lambda_1 = -8/3$, $\lambda_2 = -1/3$, $\lambda_3 = 1$.

Proposition 2.6. *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 $\sum_{i \in I} \lambda_i \overrightarrow{xa_i} = 0$ for every $x \in E$.*

Proof. By Proposition 2.5, the family $(a_i)_{i \in I}$ is affinely dependent iff the family of vectors $(\overrightarrow{a_i a_j})_{j \in (I - \{i\})}$ is linearly dependent for some $i \in I$. For any $i \in I$, the family $(\overrightarrow{a_i a_j})_{j \in (I - \{i\})}$ is linearly dependent iff there is a family $(\lambda_j)_{j \in (I - \{i\})}$ such that $\lambda_j \neq 0$ for some j, and such that

$$\sum_{j \in (I - \{i\})} \lambda_j \overrightarrow{a_i a_j} = 0.$$

Then, for any $x \in E$, we have

$$\begin{aligned} \sum_{j \in (I - \{i\})} \lambda_j \overrightarrow{a_i a_j} &= \sum_{j \in (I - \{i\})} \lambda_j (\overrightarrow{xa_j} - \overrightarrow{xa_i}) \\ &= \sum_{j \in (I - \{i\})} \lambda_j \overrightarrow{xa_j} - \left(\sum_{j \in (I - \{i\})} \lambda_j \right) \overrightarrow{xa_i}, \end{aligned}$$

and letting $\lambda_i = -\left(\sum_{j \in (I - \{i\})} \lambda_j \right)$, we get $\sum_{i \in I} \lambda_i \overrightarrow{xa_i} = 0$, with $\sum_{i \in I} \lambda_i = 0$ and $\lambda_j \neq 0$ for some $j \in I$. The converse is obvious by setting $x = a_i$ for some i such that $\lambda_i \neq 0$, since $\sum_{i \in I} \lambda_i = 0$ implies that $\lambda_j \neq 0$, for some $j \neq i$. □

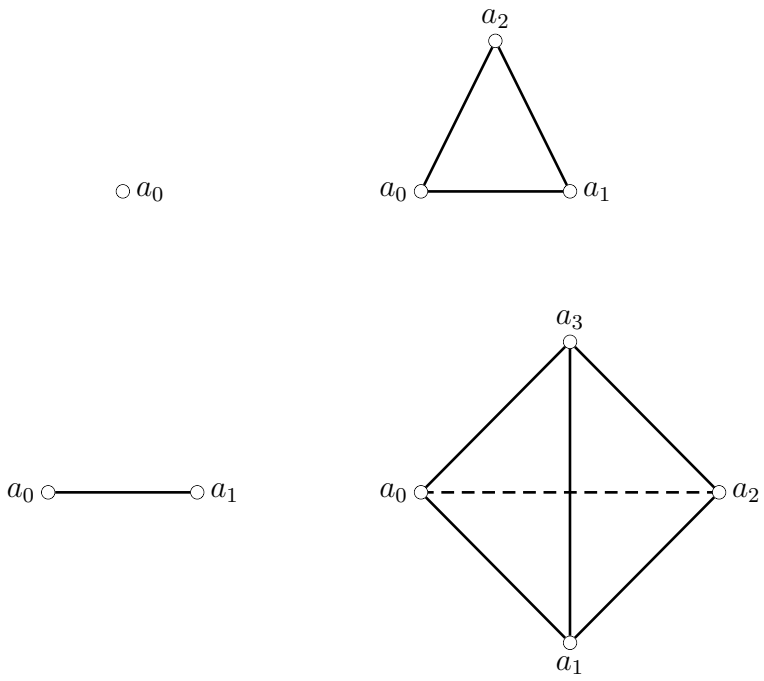


Figure 2.16: Examples of affine frames and their convex hulls.

Even though Proposition 2.6 is rather dull, it is one of the key ingredients in the proof of beautiful and deep theorems about convex sets, such as Carathéodory's theorem, Radon's theorem, and Helly's theorem.

A family of two points (a, b) in E is affinely independent iff $\overrightarrow{ab} \neq 0$, 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 $\overrightarrow{ab}$ and $\overrightarrow{ac}$ are linearly independent, which means that a, b, and c are not on the 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 $\overrightarrow{ab}$, $\overrightarrow{ac}$, and $\overrightarrow{ad}$ are linearly independent, which means that a, b, c, and d are not in the same plane (they are not coplanar). In this case, a, b, c, and d are the vertices of a tetrahedron. Figure 2.16 shows affine frames and their convex hulls for $|I| = 0, 1, 2, 3$.

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

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

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*. Figure 2.17 shows the convex hulls and associated parallelotopes for $|I| = 0, 1, 2, 3$.

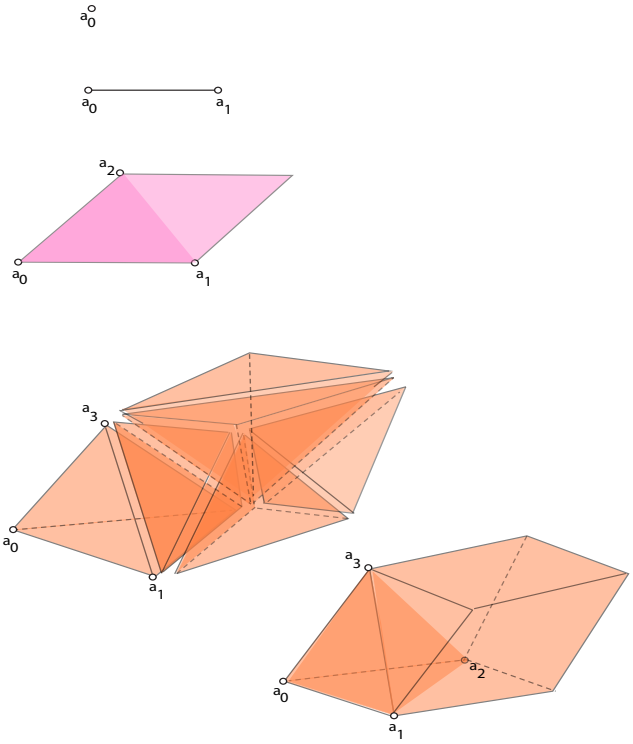


Figure 2.17: Examples of affine frames, convex hulls, and their associated parallelotopes.

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 $0 \leq \lambda \leq 1$ $(\lambda \in \mathbb{R})$.

Points are not vectors! The following example illustrates why treating points as vectors may cause problems. Let a, b, c be three affinely independent points in $\mathbb{A}^3$. Any point x in the plane (a, b, c) can be expressed as

$$x = \lambda_0 a + \lambda_1 b + \lambda_2 c,$$

where $\lambda_0 + \lambda_1 + \lambda_2 = 1$. How can we compute $\lambda_0, \lambda_1, \lambda_2$? Letting $a = (a_1, a_2, a_3)$, $b = (b_1, b_2, b_3)$, $c = (c_1, c_2, c_3)$, and $x = (x_1, x_2, x_3)$ be the coordinates of a, b, c, x in the standard frame of $\mathbb{A}^3$, it is tempting to solve the system of equations

$$\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

However, there is a problem when the origin of the coordinate system belongs to the plane (a, b, c), since in this case, the matrix is not invertible! What we should really be doing is to solve the system

$$\lambda_0\overrightarrow{Oa} + \lambda_1\overrightarrow{Ob} + \lambda_2\overrightarrow{Oc} = \overrightarrow{Ox},$$

where O is any point **not** in the plane (a, b, c). An alternative is to use certain well-chosen cross products.

It can be shown that barycentric coordinates correspond to various ratios of areas and volumes; see the problems.

2.7 Affine Maps

Corresponding to linear maps we have the notion of an affine map. An affine map is defined as a map preserving affine combinations.

Definition 2.6. Given two affine spaces $\langle E, \overrightarrow{E}, + \rangle$ and $\langle E', \overrightarrow{E'}, +' \rangle$, a function $f\colon E \to E'$ is an *affine map* iff for every family $((a_i, \lambda_i))_{i \in I}$ of weighted points in E such that $\sum_{i \in I} \lambda_i = 1$, we have

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

In other words, f preserves 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 $h\colon \overrightarrow{E} \to \overrightarrow{E'}$, we claim that the map $f\colon E \to E'$ defined such that

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

is an affine map. Indeed, for any family $(\lambda_i)_{i \in I}$ of scalars with $\sum_{i \in I} \lambda_i = 1$ and any family $(\overrightarrow{v_i})_{i \in I}$, since

$$\sum_{i \in I} \lambda_i (a + v_i) = a + \sum_{i \in I} \lambda_i \overrightarrow{a(a + v_i)} = a + \sum_{i \in I} \lambda_i v_i$$

and

$$\sum_{i \in I} \lambda_i (b + h(v_i)) = b + \sum_{i \in I} \lambda_i \overrightarrow{b(b + h(v_i))} = b + \sum_{i \in I} \lambda_i h(v_i),$$

we have

$$\begin{aligned} f\left(\sum_{i\in I}\lambda_i(a+v_i)\right) &= f\left(a+\sum_{i\in I}\lambda_i v_i\right) \\ &= b+h\left(\sum_{i\in I}\lambda_i v_i\right) \\ &= b+\sum_{i\in I}\lambda_i h(v_i) \\ &= \sum_{i\in I}\lambda_i(b+h(v_i)) \\ &= \sum_{i\in I}\lambda_i f(a+v_i). \end{aligned}$$

Note that the condition $\sum_{i\in I}\lambda_i = 1$ was implicitly used (in a hidden call to Proposition 2.1) in deriving that

$$\sum_{i\in I}\lambda_i(a+v_i) = a+\sum_{i\in I}\lambda_i v_i$$

and

$$\sum_{i\in I}\lambda_i(b+h(v_i)) = b+\sum_{i\in I}\lambda_i h(v_i).$$

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 2.18. The image of the square (a, b, c, d) is the parallelogram (a', b', c', d').

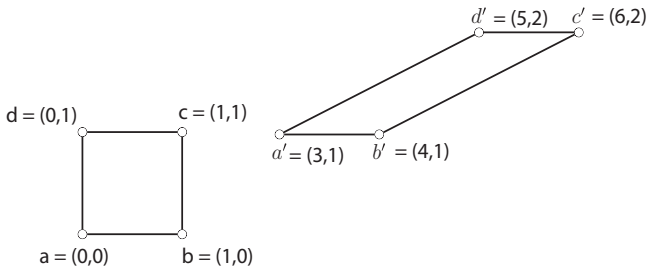


Figure 2.18: The effect of a shear.

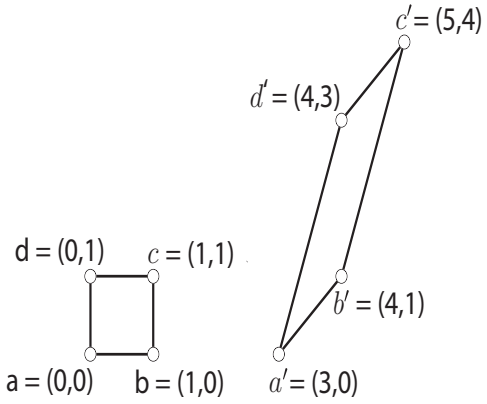


Figure 2.19: The effect of an affine map.

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

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

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 2.19. The image of the square (a, b, c, d) is the parallelogram (a', b', c', d').

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

Proposition 2.7. *Given an affine map $f\colon E \to E'$, 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 $v \in \overrightarrow{E}$.

Proof. Let $a \in E$ be any point in E. We claim that the map defined such that

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

for every $v \in \overrightarrow{E}$ is a linear map $\overrightarrow{f} : \overrightarrow{E} \to \overrightarrow{E'}$. Indeed, we can write

$$a + \lambda v = \lambda(a + v) + (1 - \lambda)a,$$

since $a + \lambda v = a + \lambda\overrightarrow{a(a+v)} + (1-\lambda)\overrightarrow{aa}$, and also

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

since $a + u + v = a + \overrightarrow{a(a+u)} + \overrightarrow{a(a+v)} - \overrightarrow{aa}$. Since f preserves barycenters, we get

$$f(a + \lambda v) = \lambda f(a + v) + (1 - \lambda)f(a).$$

If we recall that $x = \sum_{i\in I} \lambda_i a_i$ is the barycenter of a family $((a_i, \lambda_i))_{i\in I}$ of weighted points (with $\sum_{i\in I} \lambda_i = 1$) iff

$$\overrightarrow{bx} = \sum_{i\in I} \lambda_i \overrightarrow{ba_i} \quad \text{for every } b \in E,$$

we get

$$\overrightarrow{f(a)f(a+\lambda v)} = \lambda\overrightarrow{f(a)f(a+v)} + (1-\lambda)\overrightarrow{f(a)f(a)} = \lambda\overrightarrow{f(a)f(a+v)},$$

showing that $\overrightarrow{f}(\lambda v) = \lambda\overrightarrow{f}(v)$. We also have

$$f(a + u + v) = f(a + u) + f(a + v) - f(a),$$

from which we get

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

showing that $\overrightarrow{f}(u + v) = \overrightarrow{f}(u) + \overrightarrow{f}(v)$. Consequently, $\overrightarrow{f}$ is a linear map. For any other point $b \in E$, since

$$b + v = a + \overrightarrow{ab} + v = a + \overrightarrow{a(a+v)} - \overrightarrow{aa} + \overrightarrow{ab},$$

$b + v = (a + v) - a + b$, and since f preserves barycenters, we get

$$f(b + v) = f(a + v) - f(a) + f(b),$$

which implies that

$$\begin{aligned}
\overrightarrow{f(b)f(b+v)} &= \overrightarrow{f(b)f(a+v)} - \overrightarrow{f(b)f(a)} + \overrightarrow{f(b)f(b)}, \\
&= \overrightarrow{f(a)f(b)} + \overrightarrow{f(b)f(a+v)}, \\
&= \overrightarrow{f(a)f(a+v)}.
\end{aligned}$$

Thus, $\overrightarrow{f(b)f(b+v)} = \overrightarrow{f(a)f(a+v)}$, which shows that the definition of $\overrightarrow{f}$ does not depend on the choice of $a \in E$. The fact that $\overrightarrow{f}$ is unique is obvious: We must have $\overrightarrow{f}(v) = \overrightarrow{f(a)f(a+v)}$. □

The unique linear map $\overrightarrow{f} : \overrightarrow{E} \to \overrightarrow{E'}$ given by Proposition 2.7 is called 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

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

for all $a, x \in E$. Proposition 2.7 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 $\overrightarrow{f} : \overrightarrow{E} \to \overrightarrow{E'}$, such that

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

for all $v \in \overrightarrow{E}$ (just let $b = f(a)$, for any $a \in E$). Affine maps for which $\overrightarrow{f}$ is the identity map are called *translations*. Indeed, if $\overrightarrow{f} = \mathrm{id}$,

$$\begin{aligned} f(x) &= f(a) + \overrightarrow{f}(\overrightarrow{ax}) = f(a) + \overrightarrow{ax} = x + \overrightarrow{xa} + \overrightarrow{af(a)} + \overrightarrow{ax} \\ &= x + \overrightarrow{xa} + \overrightarrow{af(a)} - \overrightarrow{xa} = x + \overrightarrow{af(a)}, \end{aligned}$$

and so

$$\overrightarrow{xf(x)} = \overrightarrow{af(a)},$$

which shows that f is the translation induced by the vector $\overrightarrow{af(a)}$ (which does not depend on a).

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, and the image of a plane is either a point, a line, or a plane.

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

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

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, then for any other affine space F and for any sequence $(b_0, b_1, \ldots, b_m)$ of $m + 1$ points in F, there

is a unique affine map $f\colon E \to F$ such that $f(a_i) = b_i$, for $0 \leq i \leq m$. Indeed, f must be such that

$$f(\lambda_0 a_0 + \cdots + \lambda_m a_m) = \lambda_0 b_0 + \cdots + \lambda_m b_m,$$

where $\lambda_0 + \cdots + \lambda_m = 1$, and this defines a unique affine map on all of E, since $(a_0, a_1, \ldots, a_m)$ is an affine frame for E.

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, the more general case where an affine map $f\colon E \to F$ is represented with respect to two affine frames $(a_0, \ldots, a_n)$ in E and $(b_0, \ldots, b_m)$ in F being analogous. Since

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

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

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

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

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

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

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

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

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

can also be written as

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

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 $\frac{1}{2}$ 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

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

has no fixed point, even though

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

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 $f\colon E \to E'$ 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 $f\colon E \to E'$ 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 [7] or Samuel [52]).

Given three collinear points 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

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

provided that $\beta \neq 1$, i.e., $b \neq c$. 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 $-\text{ratio}(a, b, c) = \frac{\overrightarrow{ba}}{\overrightarrow{bc}}$. Since affine maps preserve barycenters, it is clear that affine maps preserve the ratio of three points.

2.8 Affine Groups

We now take a quick look at the bijective affine maps. Given an affine space E, the set of affine bijections $f\colon E \to E$ is clearly a group, called the *affine group of* E, and denoted by $\mathbf{GA}(E)$. Recall that the group of bijective linear maps of the vector space $\overrightarrow{E}$ is denoted by $\mathbf{GL}(\overrightarrow{E})$. Then, the map $f \mapsto \overrightarrow{f}$ defines a group homomorphism $L\colon \mathbf{GA}(E) \to \mathbf{GL}(\overrightarrow{E})$. 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 $\mathbf{GL}(\overrightarrow{E})$, and is denoted by $\mathbb{R}^*\mathrm{id}_{\overrightarrow{E}}$ (where $\lambda\, \mathrm{id}_{\overrightarrow{E}}(u) = \lambda u$, and $\mathbb{R}^* = \mathbb{R} - \{0\}$). The subgroup $\mathbf{DIL}(E) = L^{-1}(\mathbb{R}^*\mathrm{id}_{\overrightarrow{E}})$ of $\mathbf{GA}(E)$ 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 $\mathbf{DIL}(E)$ are called *affine dilatations*.

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

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

for every $x \in E$.

Remark: The terminology does not seem to be universally agreed upon. The terms *affine dilatation* and *central dilatation* are used by Pedoe [48]. Snapper and Troyer use the term *dilation* for an affine dilatation and *magnification* for a central dilatation [55]. Samuel uses *homothety* for a central dilatation, a direct translation of the French "homothétie" [52]. Since dilation is shorter than dilatation and somewhat easier to pronounce, perhaps we should use that!

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" $\overrightarrow{ax}$ by λ.

Figure 2.20 shows the effect of a central dilatation of center d. The triangle (a, b, c) is magnified to the triangle (a', b', c'). Note how every line is mapped to a parallel line.

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.

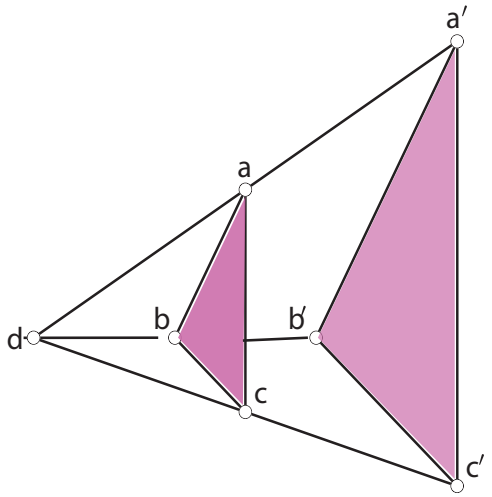


Figure 2.20: The effect of a central dilatation $H_{d,\lambda}(x)$.

Proposition 2.8. *Given any affine space E, for any affine bijection $f \in \mathbf{GA}(E)$, 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 $f = H_{c,\lambda}$.*

Proof. The proof is straightforward, and is omitted. It is also given in Gallier [29]. □

Clearly, if $\overrightarrow{f} = \mathrm{id}_{\overrightarrow{E}}$, the affine map f is a translation. Thus, the group of affine dilatations $\mathbf{DIL}(E)$ 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.

Another point worth mentioning is that affine bijections preserve the ratio of volumes of parallelotopes. Indeed, given any basis $B = (u_1, \ldots, u_m)$ of the vector space $\overrightarrow{E}$ associated with the affine space E, given any $m+1$ affinely independent points $(a_0, \ldots, a_m)$, we can compute the determinant $\det_B(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ w.r.t. the basis B. For any bijective affine map $f\colon E \to E$, since

$$\det_B\Big(\overrightarrow{f}(\overrightarrow{a_0a_1}), \ldots, \overrightarrow{f}(\overrightarrow{a_0a_m})\Big) = \det\big(\overrightarrow{f}\big)\det_B(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$$

and the determinant of a linear map is intrinsic (i.e., depends only on $\overrightarrow{f}$, and not on the particular basis B), we conclude that the ratio

$$\frac{\det_B\Big(\overrightarrow{f}(\overrightarrow{a_0a_1}), \ldots, \overrightarrow{f}(\overrightarrow{a_0a_m})\Big)}{\det_B(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})} = \det\big(\overrightarrow{f}\big)$$

is independent of the basis B. Since $\det_B(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ is the volume of the parallelotope spanned by $(a_0, \ldots, a_m)$, where the parallelotope spanned by any point a and the vectors

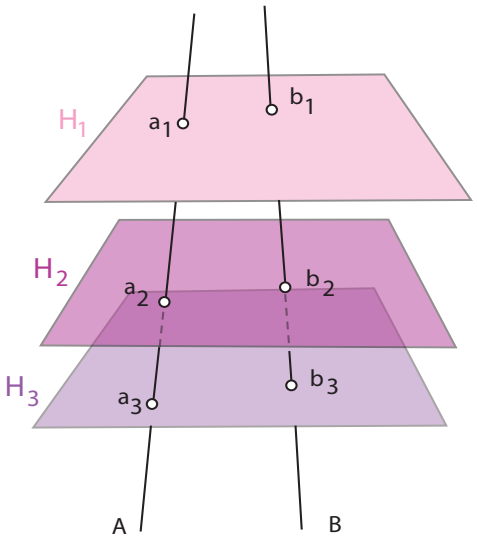


Figure 2.21: The theorem of Thales.

$(u_1, \ldots, u_m)$ has unit volume (see Berger [7], Section 9.12), we see that affine bijections preserve the ratio of volumes of parallelotopes. In fact, this ratio is independent of the choice of the parallelotopes of unit volume. In particular, the affine bijections $f \in \mathbf{GA}(E)$ such that $\det(\overrightarrow{f}) = 1$ preserve volumes. These affine maps form a subgroup $\mathbf{SA}(E)$ of $\mathbf{GA}(E)$ called the *special affine group of* E. We now take a glimpse at affine geometry.

2.9 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 Thales.

Proposition 2.9. *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:*

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

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

Proof. Figure 2.21 illustrates the theorem of Thales. We sketch a proof, leaving the details as an exercise. Since H_1, H_2, H_3 are parallel, they have the same direction $\overrightarrow{H}$, a hyperplane

in $\overrightarrow{E}$. Let $u \in \overrightarrow{E} - \overrightarrow{H}$ be any nonnull vector such that $A = a_1 + \mathbb{R}u$. Since A is not parallel to H, we have $\overrightarrow{E} = \overrightarrow{H} \oplus \mathbb{R}u$, and thus we can define the linear map $p\colon \overrightarrow{E} \to \mathbb{R}u$, the projection on $\mathbb{R}u$ parallel to $\overrightarrow{H}$. This linear map induces an affine map $f\colon E \to A$, by defining f such that

$$f(b_1 + w) = a_1 + p(w),$$

for all $w \in \overrightarrow{E}$. Clearly, $f(b_1) = a_1$, and since H_1, H_2, H_3 all have direction $\overrightarrow{H}$, we also have $f(b_2) = a_2$ and $f(b_3) = a_3$. Since f is affine, it preserves ratios, and thus

$$\frac{\overrightarrow{a_1a_3}}{\overrightarrow{a_1a_2}} = \frac{\overrightarrow{b_1b_3}}{\overrightarrow{b_1b_2}}.$$

The converse is immediate. □

We also have the following simple proposition, whose proof is left as an easy exercise.

Proposition 2.10. *Given any affine space E, given any two distinct points $a, b \in E$, and 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 $b' = D' \cap \langle c, b\rangle$.

The first case is the parallelogram law, and the second case follows easily from Thales' theorem. For an illustration, see Figure 2.22.

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

Proposition 2.11. *Given any affine plane E, any two distinct lines D and D', then 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.*

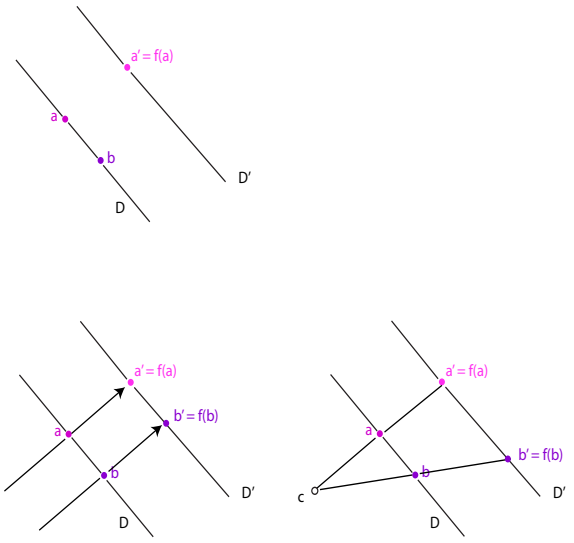


Figure 2.22: An illustration of Proposition 2.10. The bottom left diagram illustrates a translation, while the bottom right illustrates a central dilation through c.

Proof. Pappus's theorem is illustrated in Figure 2.23. If D and D' are not parallel, let d be their intersection. Let f be the dilatation of center d such that $f(a) = b$, and let g be the dilatation of center d such that $g(b) = c$. Since 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, by Proposition 2.10 we have $a' = f(b')$ and $b' = g(c')$. However, we observed that dilatations with the same center commute, and thus $f \circ g = g \circ f$, and thus, letting $h = g \circ f$, we get $c = h(a)$ and $a' = h(c')$. Again, by Proposition 2.10, the lines $\langle a, c' \rangle$ and $\langle a', c \rangle$ are parallel. If D and D' are parallel, we use translations instead of dilatations. □

There is a converse to Pappus's theorem, which yields a fancier version of Pappus's theorem, but it is easier to prove it using projective geometry. It should be noted that in axiomatic presentations of projective geometry, Pappus's theorem is equivalent to the commutativity of the ground field K (in the present case, $K = \mathbb{R}$). We now prove an affine version of Desargues's theorem.

Proposition 2.12. *Given any affine space E, and 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).*

Proof. We prove half of the proposition, the direction in which it is assumed that $\langle a, c \rangle$ and $\langle a', c' \rangle$ are parallel, leaving the converse as an exercise. Since the lines $\langle a, b \rangle$ and $\langle a', b' \rangle$ are

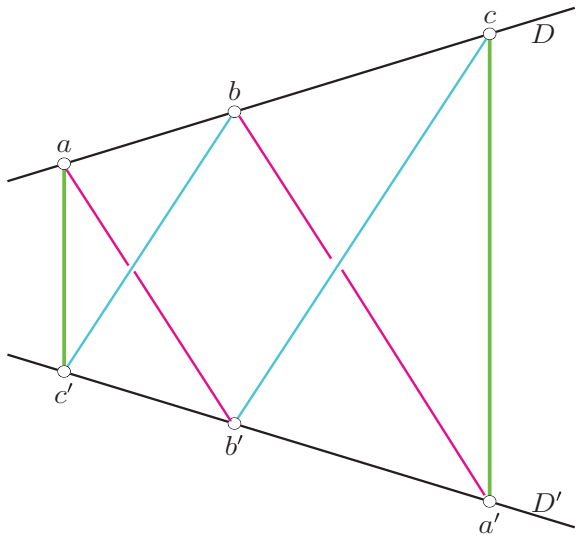


Figure 2.23: Pappus's theorem (affine version).

parallel, the points a, b, a', b' are coplanar. Thus, either $\langle a, a'\rangle$ and $\langle b, b'\rangle$ are parallel, or they have some intersection d. We consider the second case where they intersect, leaving the other case as an easy exercise. Let f be the dilatation of center d such that $f(a) = a'$. By Proposition 2.10, we get $f(b) = b'$. If $f(c) = c''$, again by Proposition 2.10 twice, the lines $\langle b, c\rangle$ and $\langle b', c''\rangle$ are parallel, and the lines $\langle a, c\rangle$ and $\langle a', c''\rangle$ are parallel. From this it follows that $c'' = c'$. Indeed, recall that $\langle b, c\rangle$ and $\langle b', c'\rangle$ are parallel, and similarly $\langle a, c\rangle$ and $\langle a', c'\rangle$ are parallel. Thus, the lines $\langle b', c''\rangle$ and $\langle b', c'\rangle$ are identical, and similarly the lines $\langle a', c''\rangle$ and $\langle a', c'\rangle$ are identical. Since $\overrightarrow{a'c'}$ and $\overrightarrow{b'c'}$ are linearly independent, these lines have a unique intersection, which must be $c'' = c'$.

The direction where it is assumed that the lines $\langle a, a'\rangle$, $\langle b, b'\rangle$ and $\langle c, c'\rangle$, are either parallel or concurrent is left as an exercise (in fact, the proof is quite similar). □

Desargues's theorem is illustrated in Figure 2.24.

There is a fancier version of Desargues's theorem, but it is easier to prove it using projective geometry. It should be noted that in axiomatic presentations of projective geometry, Desargues's theorem is related to the associativity of the ground field K (in the present case, $K = \mathbb{R}$). Also, Desargues's 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. We leave the proof as an exercise.

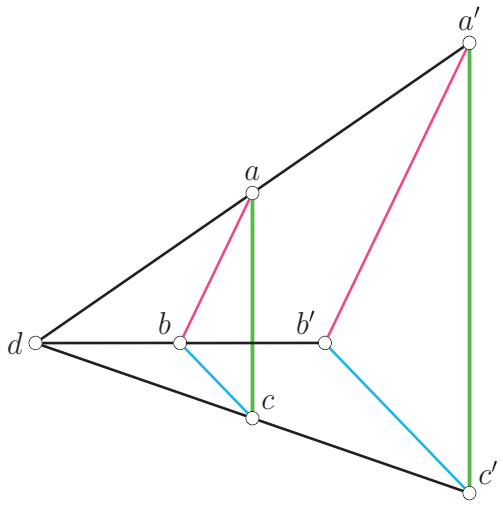


Figure 2.24: Desargues's theorem (affine version).

2.10 Affine Hyperplanes

We now consider affine forms and affine hyperplanes. In Section 2.5 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. Given an affine map $f\colon E \to \mathbb{R}$, we also denote $f^{-1}(0)$ by $\operatorname{Ker} f$, and we call it the *kernel* of f. Recall that an (affine) hyperplane is an affine subspace of codimension 1. The relationship between affine hyperplanes and affine forms is given by the following proposition.

Proposition 2.13. *Let E be an affine space. The following properties hold:*

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

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

Proof. The proof is straightforward, and is omitted. It is also given in Gallier [29]. □

When E is of dimension n, given an affine frame $(a_0, (u_1, \ldots, u_n))$ of E with origin a_0, recall from Definition 2.5 that every point of E can be expressed uniquely as $x = a_0 + x_1u_1 + \cdots + x_nu_n$, where $(x_1, \ldots, x_n)$ are the *coordinates* of x with respect to the affine frame $(a_0, (u_1, \ldots, u_n))$.

Also recall that every linear form f^* is such that $f^*(x) = \lambda_1 x_1 + \cdots + \lambda_n x_n$, for every $x = x_1u_1 + \cdots + x_nu_n$ and some $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Since an affine form $f\colon E \to \mathbb{R}$ satisfies the property $f(a_0 + x) = f(a_0) + \overrightarrow{f}(x)$, denoting $f(a_0 + x)$ by $f(x_1, \ldots, x_n)$, we see that we have

$$f(x_1, \ldots, x_n) = \lambda_1 x_1 + \cdots + \lambda_n x_n + \mu,$$

where $\mu = f(a_0) \in \mathbb{R}$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Thus, a hyperplane is the set of points whose coordinates $(x_1, \ldots, x_n)$ satisfy the (affine) equation

$$\lambda_1 x_1 + \cdots + \lambda_n x_n + \mu = 0.$$

2.11 Intersection of Affine Spaces

In this section we take a closer look at the intersection of affine subspaces. This subsection can be omitted at first reading.

First, we need a result of linear algebra. Given a vector space E and any two subspaces M and N, there are several interesting linear maps. We have the canonical injections $i\colon M \to M+N$ and $j\colon N \to M+N$, the canonical injections $in_1\colon M \to M \oplus N$ and $in_2\colon N \to M \oplus N$, and thus, injections $f\colon M \cap N \to M \oplus N$ and $g\colon M \cap N \to M \oplus N$, where f is the composition of the inclusion map from $M \cap N$ to M with in_1, and g is the composition of the inclusion map from $M \cap N$ to N with in_2. Then, we have the maps $f+g\colon M \cap N \to M \oplus N$, and $i-j\colon M \oplus N \to M+N$.

Proposition 2.14. *Given a vector space E and any two subspaces M and N, with the definitions above,*

$$0 \longrightarrow M \cap N \xrightarrow{f+g} M \oplus N \xrightarrow{i-j} M+N \longrightarrow 0$$

is a short exact sequence, which means that $f+g$ is injective, $i-j$ is surjective, and that $\mathrm{Im}\,(f+g) = \mathrm{Ker}\,(i-j)$. As a consequence, we have the Grassmann relation

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

Proof. It is obvious that $i-j$ is surjective and that $f+g$ is injective. Assume that $(i-j)(u+v) = 0$, where $u \in M$, and $v \in N$. Then, $i(u) = j(v)$, and thus, by definition of i and j, there is some $w \in M \cap N$, such that $i(u) = j(v) = w \in M \cap N$. By definition of f and g, $u = f(w)$ and $v = g(w)$, and thus $\mathrm{Im}\,(f+g) = \mathrm{Ker}\,(i-j)$, as desired. The second part of the proposition follows from standard results of linear algebra (see Artin [3], Strang [60], or Lang [38]). □

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

Proposition 2.15. *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 $\overrightarrow{ab} \in \overrightarrow{M} + \overrightarrow{N}$ for some $a \in M$ and some $b \in N$.

(2) $M \cap N$ consists of a single point iff $\overrightarrow{ab} \in \overrightarrow{M} + \overrightarrow{N}$ 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 $\overrightarrow{S} = \overrightarrow{M} + \overrightarrow{N} + K\overrightarrow{ab}$ (the vector space $\overrightarrow{E}$ is defined over the field K).

Proof. (1) Pick any $a \in M$ and any $b \in N$, which is possible, since M and N are nonempty. Since $\overrightarrow{M} = \{\overrightarrow{ax} \mid x \in M\}$ and $\overrightarrow{N} = \{\overrightarrow{by} \mid y \in N\}$, if $M \cap N \neq \emptyset$, for any $c \in M \cap N$ we have $\overrightarrow{ab} = \overrightarrow{ac} - \overrightarrow{bc}$, with $\overrightarrow{ac} \in \overrightarrow{M}$ and $\overrightarrow{bc} \in \overrightarrow{N}$, and thus, $\overrightarrow{ab} \in \overrightarrow{M} + \overrightarrow{N}$. Conversely, assume that $\overrightarrow{ab} \in \overrightarrow{M} + \overrightarrow{N}$ for some $a \in M$ and some $b \in N$. Then $\overrightarrow{ab} = \overrightarrow{ax} + \overrightarrow{by}$, for some $x \in M$ and some $y \in N$. But we also have

$$\overrightarrow{ab} = \overrightarrow{ax} + \overrightarrow{xy} + \overrightarrow{yb},$$

and thus we get $0 = \overrightarrow{xy} + \overrightarrow{yb} - \overrightarrow{by}$, that is, $\overrightarrow{xy} = 2\overrightarrow{by}$. Thus, b is the middle of the segment $[x, y]$, and since $\overrightarrow{yx} = 2\overrightarrow{yb}$, $x = 2b - y$ is the barycenter of the weighted points $(b, 2)$ and $(y, -1)$. Thus x also belongs to N, since N being an affine subspace, it is closed under barycenters. Thus, $x \in M \cap N$, and $M \cap N \neq \emptyset$.

(2) Note that in general, if $M \cap N \neq \emptyset$, then

$$\overrightarrow{M \cap N} = \overrightarrow{M} \cap \overrightarrow{N},$$

because

$$\overrightarrow{M \cap N} = \{\overrightarrow{ab} \mid a, b \in M \cap N\} = \{\overrightarrow{ab} \mid a, b \in M\} \cap \{\overrightarrow{ab} \mid a, b \in N\} = \overrightarrow{M} \cap \overrightarrow{N}.$$

Since $M \cap N = c + \overrightarrow{M \cap N}$ for any $c \in M \cap N$, we have

$$M \cap N = c + \overrightarrow{M} \cap \overrightarrow{N} \quad \text{for any } c \in M \cap N.$$

From this it follows that if $M \cap N \neq \emptyset$, then $M \cap N$ consists of a single point iff $\overrightarrow{M} \cap \overrightarrow{N} = \{0\}$. This fact together with what we proved in (1) proves (2).

(3) This is left as an easy exercise. □

Remarks:

(1) The proof of Proposition 2.15 shows that if $M \cap N \neq \emptyset$, then $\overrightarrow{ab} \in \overrightarrow{M} + \overrightarrow{N}$ for all $a \in M$ and all $b \in N$.

(2) Proposition 2.15 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. Indeed, if $\overrightarrow{E} = \overrightarrow{M} \oplus \overrightarrow{N}$, then $\overrightarrow{ab} \in \overrightarrow{E}$ for all $a \in M$ and all $b \in N$, and since $\overrightarrow{M} \cap \overrightarrow{N} = \{0\}$, the result follows from part (2) of the proposition.

We can now state the following proposition.

Proposition 2.16. *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

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

Proof. The proof is not difficult, using Proposition 2.15 and Proposition 2.14, but we leave it as an exercise. □

Chapter 3

Basic Properties of Convex Sets

Convex sets play a very important role in geometry. In this chapter we state and prove 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.

3.1 A Review of Basic Topological Concepts

Given an affine space E with associated vector space $\overrightarrow{E}$, for any two points $a, b \in E$, the unique vector from a to b is denoted $\mathbf{ab}$ rather than $\overrightarrow{ab}$, so that $b = a + \mathbf{ab}$.

The vector space $\mathbb{R}^d$ viewed as an affine space is denoted by $\mathbb{A}^d$. In addition, if $\mathbb{R}^d$ is equipped with the standard Euclidean inner product and $\mathbb{R}^d$ is viewed a an affine space, then it is denoted by $\mathbb{E}^d$.

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

3.2 Convex Sets

Convex sets are defined as follows.

Definition 3.1. A subset V of a real affine space 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 $0 \leq \lambda \leq 1$ ($\lambda \in \mathbb{R}$). Given any two points a, b, the notation $[a, b]$ is often used to denote the line segment between a and b, that is,

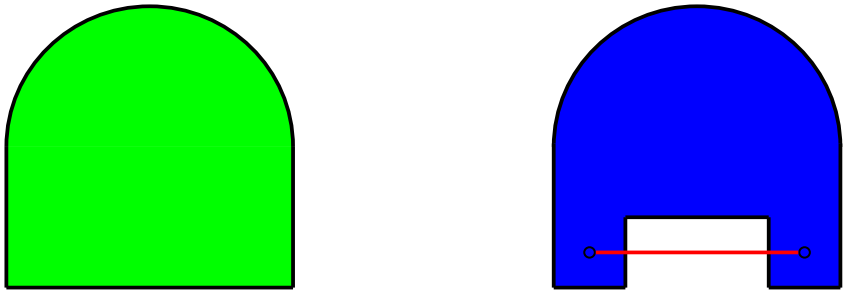
$$[a, b] = \{c \in E \mid c = (1-\lambda)a + \lambda b,\ 0 \leq \lambda \leq 1\},$$

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 by $\operatorname{conv}(S)$ or $\mathcal{C}(S)$, and called the *convex hull of* S (namely, the intersection of all convex sets containing S). The *affine hull* of a subset S of E is the smallest affine set containing S and it will be denoted by $\langle S\rangle$ or $\operatorname{aff}(S)$.

Definition 3.2. Given any affine space E, the *dimension* of a nonempty convex subset S of E, denoted by $\dim S$, is the dimension of the smallest affine subset $\operatorname{aff}(S)$ containing S.



(a) (b)

Figure 3.1: (a) A convex set; (b) A nonconvex set

A good understanding of what $\mathrm{conv}(S)$ is, and good methods for computing it, are essential. First we have the following simple but crucial lemma:

Lemma 3.1. *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 $(a_i)_{i \in I}$.*

Proof. If $(a_i)_{i \in I}$ is empty, then $V = \emptyset$, because of the condition $\sum_{i \in I} \lambda_i = 1$. As in the case of affine combinations, it is easily shown by induction that any convex combination can be obtained by computing convex combinations of two points at a time. As a consequence, if $(a_i)_{i \in I}$ is nonempty, then the smallest convex subspace containing $(a_i)_{i \in I}$ must contain the set V of all convex combinations $\sum_{i \in I} \lambda_i a_i$. Thus, it is enough to show that V is closed under convex combinations, which is immediately verified. □

In view of Lemma 3.1, 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., $H = \mathrm{Ker}\, f$, with $f(a + u) = f(a) + \overrightarrow{f}(u)$ for all $a \in E$ and all $u \in \overrightarrow{E}$, where $f(a) \in \mathbb{R}$ and $\overrightarrow{f} : \overrightarrow{E} \to \mathbb{R}$ is a nonzero linear form), we can define the two subsets

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

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

$$\{H_+(f), H_-(f)\}$$

depends only 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 $H_+(f)$

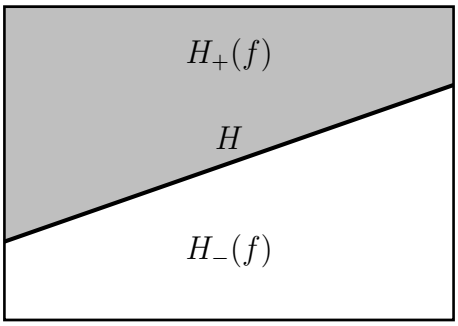


Figure 3.2: The two half-spaces determined by a hyperplane, H

and $H_-(f)$ are the *closed half-spaces associated with* H. Clearly, $H_+(f) \cup H_-(f) = E$ and $H_+(f) \cap H_-(f) = H$. 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*).

The convex combinations $\sum_{i\in I} \lambda_i a_i$ arising in computing the convex hull of a family of points $(a_i)_{i\in I}$ have finite support, so they can be written as $\sum_{j\in J} \lambda_j a_j$ for some *finite* subset J of I, but a priori there is no bound on the size such finite sets J. Thus it is natural to wonder whether Lemma 3.1 can be sharpened in two directions:

(1) Is it possible to have a fixed bound on the number of points involved in the convex combinations $\sum_{j\in J} \lambda_j a_j$ (that is, on the size of the index sets J)?

(2) Is it necessary to consider convex combinations of all points, or is it possible to consider only a subset of points 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. For example, in the plane $\mathbb{A}^2$, the convex hull of a set S of points is the union of all triangles (interior points included) with vertices in S. In Case (2), the theorem of Krein and Milman asserts that a convex set that is also compact is the convex hull of its extremal points (given a convex set S, a point $a \in S$ is extremal if $S - \{a\}$ is also convex, see Berger [8] or Lang [39]). Next, we prove Carathéodory's theorem.

3.3 Carathéodory's Theorem

The proof of Carathéodory's theorem is really beautiful. It proceeds by contradiction and uses a minimality argument.

Theorem 3.2. *(Carathéodory, 1907) Given any affine space E of dimension m, for any (nonvoid) family $S = (a_i)_{i\in L}$ in E, the convex hull* $\mathrm{conv}(S)$ *of S is equal to the set of convex combinations of families of $m+1$ points of S.*

Proof. By Lemma 3.1,

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

We would like to prove that

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

We proceed by contradiction. If the theorem is false, there is some point $b \in \mathrm{conv}(S)$ 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 fewer than q points in S, that is, where $|J| < q$. Such a point $b \in \mathrm{conv}(S)$ is a convex combination

$$b = \lambda_1 a_1 + \cdots + \lambda_q a_q,$$

where $\lambda_1 + \cdots + \lambda_q = 1$ and $\lambda_i > 0$ $(1 \leq i \leq q)$. We shall prove that b can be written as a convex combination of $q-1$ of the a_i. Pick any origin O in E. Since there are $q > m+1$ points $a_1, \ldots, a_q$, these points are affinely dependent, and by Lemma 2.6.5 from Gallier [30], there is a family $(\mu_1, \ldots, \mu_q)$ all scalars not all null, such that $\mu_1 + \cdots + \mu_q = 0$ and

$$\sum_{i=1}^{q} \mu_i \mathbf{Oa_i} = 0.$$

Consider the set $T \subseteq \mathbb{R}$ defined by

$$T = \{t \in \mathbb{R} \mid \lambda_i + t\mu_i \geq 0,\ \mu_i \neq 0,\ 1 \leq i \leq q\}.$$

The set T is nonempty, since it contains 0. Since $\sum_{i=1}^{q} \mu_i = 0$ and the μ_i are not all null, there are some μ_h, μ_k such that $\mu_h < 0$ and $\mu_k > 0$, which implies that $T = [\alpha, \beta]$, where

$$\alpha = \max_{1\leq i\leq q} \{-\lambda_i/\mu_i \mid \mu_i > 0\} \quad \text{and} \quad \beta = \min_{1\leq i\leq q} \{-\lambda_i/\mu_i \mid \mu_i < 0\}$$

(T is the intersection of the closed half-spaces $\{t \in \mathbb{R} \mid \lambda_i + t\mu_i \geq 0,\ \mu_i \neq 0\}$). Observe that $\alpha < 0 < \beta$, since $\lambda_i > 0$ for all $i = 1, \ldots, q$.

We claim that there is some j $(1 \leq j \leq q)$ such that

$$\lambda_j + \alpha\mu_j = 0.$$

Indeed, since

$$\alpha = \max_{1 \leq i \leq q} \{-\lambda_i/\mu_i \mid \mu_i > 0\},$$

as the set on the right hand side is finite, the maximum is achieved and there is some index j so that $\alpha = -\lambda_j/\mu_j$. If j is some index such that $\lambda_j + \alpha\mu_j = 0$, since $\sum_{i=1}^{q} \mu_i \mathbf{Oa_i} = 0$, we have

$$\begin{aligned}
b &= \sum_{i=1}^{q} \lambda_i a_i = O + \sum_{i=1}^{q} \lambda_i \mathbf{Oa_i} + 0, \\
&= O + \sum_{i=1}^{q} \lambda_i \mathbf{Oa_i} + \alpha \left(\sum_{i=1}^{q} \mu_i \mathbf{Oa_i} \right), \\
&= O + \sum_{i=1}^{q} (\lambda_i + \alpha\mu_i) \mathbf{Oa_i}, \\
&= \sum_{i=1}^{q} (\lambda_i + \alpha\mu_i) a_i, \\
&= \sum_{i=1,\, i \neq j}^{q} (\lambda_i + \alpha\mu_i) a_i,
\end{aligned}$$

since $\lambda_j + \alpha\mu_j = 0$. Since $\sum_{i=1}^{q} \mu_i = 0$, $\sum_{i=1}^{q} \lambda_i = 1$, and $\lambda_j + \alpha\mu_j = 0$, we have

$$\sum_{i=1,\, i \neq j}^{q} \lambda_i + \alpha\mu_i = 1,$$

and since $\lambda_i + \alpha\mu_i \geq 0$ for $i = 1, \ldots, q$, the above shows that b can be expressed as a convex combination of $q - 1$ points from S. However, this contradicts the assumption that b cannot be expressed as a convex combination of strictly fewer than q points from S, and the theorem is proved. □

If S is a finite (of infinite) set of points in the affine plane $\mathbb{A}^2$, Theorem 3.2 confirms our intuition that $\mathrm{conv}(S)$ 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!

An interesting consequence of Carathéodory's theorem is the following result:

Proposition 3.3. *If K is any compact subset of $\mathbb{A}^m$, then the convex hull,* $\mathrm{conv}(K)$*, of K is also compact. In particular, the convex hull* $\mathrm{conv}(a_1, \ldots, a_p)$ *of a finite set of points is compact, and thus closed (and bounded).*

Proof. Let C be the subset of $\mathbb{A}^{m+1}$ given by

$$C = \{(\lambda_0, \ldots, \lambda_m) \in \mathbb{R}^{m+1} \mid \lambda_0 + \cdots + \lambda_m = 1,\ \lambda_i \geq 0,\ i = 0, \ldots, m\}.$$

Clearly C is cut out by hyperplanes and is bounded (since $0 \leq \lambda_i \leq 1$) so C is compact (and also convex), and as K is compact, $C \times K^{m+1}$ is also compact. Then, consider the map $f\colon C \times (\mathbb{A}^m)^{m+1} \to \mathbb{A}^m$ given by

$$f(\lambda_0, \ldots, \lambda_m, a_0, \ldots, a_m) = \lambda_0 a_0 + \cdots + \lambda_m a_m.$$

This is obviously a continuous map. Furthermore, by Carathéodory's theorem,

$$f(C \times K^{m+1}) = \mathrm{conv}(K),$$

and since the image of a compact set by a continuous function is compact, we conclude that $\mathrm{conv}(K)$ is compact. □

A closer examination of the proof of Theorem 3.2 reveals that the fact that the μ_i's add up to zero ensures that T is a closed interval, but all we need is that T be bounded from below, and this only requires that some μ_j be strictly positive. As a consequence, we can prove a version of Theorem 3.2 for convex cones. This is a useful result since cones play such an important role in convex optimization. let us recall some basic definitions about cones.

Definition 3.3. Given any vector space E, a subset $C \subseteq E$ is a *convex cone* iff C is closed under *positive linear combinations*, that is, linear combinations of the form

$$\sum_{i \in I} \lambda_i v_i, \quad \text{with} \quad v_i \in C \quad \text{and} \quad \lambda_i \geq 0 \quad \text{for all} \quad i \in I,$$

where I has finite support (all $\lambda_i = 0$ except for finitely many $i \in I$). Given any set of vectors S, the *positive hull* of S, or *cone* spanned by S, denoted $\mathrm{cone}(S)$, is the set of all positive linear combinations of vectors in S,

$$\mathrm{cone}(S) = \left\{ \sum_{i \in I} \lambda_i v_i \mid v_i \in S,\ \lambda_i \geq 0 \right\}.$$

Note that a cone always contains 0. When S consists of a finite number of vector, the convex cone $\mathrm{cone}(S)$ is called a *polyhedral cone*. We have the following version of Carathéodory's theorem for convex cones:

Theorem 3.4. *Given any vector space E of dimension m, for any (nonvoid) family $S = (v_i)_{i \in L}$ of vectors in E, the cone* $\mathrm{cone}(S)$ *spanned by S is equal to the set of positive combinations of families of m vectors in S.*

The proof of Theorem 3.4 can be easily adapted from the proof of Theorem 3.2 and is left as an exercise.

There is an interesting generalization of Carathéodory's theorem known as the *Colorful Carathéodory theorem*. This theorem due to Bárány and proved in 1982 can be used to give a fairly short proof of a generalization of Helly's theorem known as Tverberg's theorem (see Section 3.5).

Theorem 3.5. *(Colorful Carathéodory theorem) Let E be any affine space of dimension m. For any point $b \in E$, for any sequence of $m+1$ nonempty subsets $(S_1, \ldots, S_{m+1})$ of E, if $b \in \text{conv}(S_i)$ for $i = 1, \ldots, m+1$, then there exists a sequence of $m+1$ points $(a_1, \ldots, a_{m+1})$, with $a_i \in S_i$, so that $b \in \text{conv}(a_1, \ldots, a_{m+1})$, that is, b is a convex combination of the a_i's.*

Although Theorem 3.5 is not hard to prove, we will not prove it here. Instead, we refer the reader to Matousek [41], Chapter 8, Section 8.2. There is also a stronger version of Theorem 3.5, in which it is enough to assume that $b \in \text{conv}(S_i \cup S_j)$ for all i, j with $1 \leq i < j \leq m+1$.

Now that we have given an answer to the first question posed at the end of Section 3.2 we give an answer to the second question.

3.4 Vertices, Extremal Points and Krein and Milman's Theorem

First, we define the notions of separation and of separating hyperplanes. 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., $H = \text{Ker}\, f$), 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 $\{H_+(f), H_-(f)\}$ depends only 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, we have $\overset{\circ}{H}_+(f) = H_+(f) - H$ and $\overset{\circ}{H}_-(f) = H_-(f) - H$.

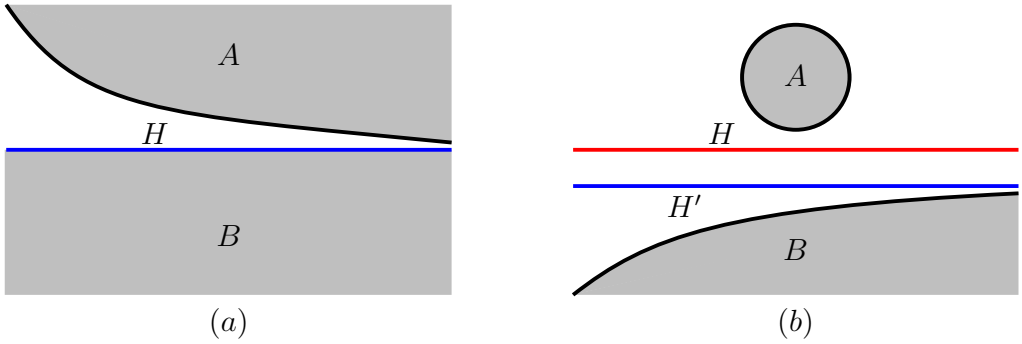


Figure 3.3: (a) A separating hyperplane H. (b) Strictly separating hyperplanes H and H'.

Definition 3.4. Given an affine space E and two nonempty subsets A and B of E, 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.

In Figure 3.3 (a), the two closed convex sets A and B are unbounded and B has the hyperplane H for its boundary, while A is asymptotic to H. The hyperplane H is a separating hyperplane for A and B but A and B can't be strictly separated. In Figure 3.3 (b), both A and B are convex and closed, B is unbounded and asymptotic to the hyperplane, H', but A is bounded. Both hyperplanes H and H' strictly separate A and B.

The special case of separation where A is convex and $B = \{a\}$, for some point, a, in A, is of particular importance.

Definition 3.5. Let E be an affine space and let A be any nonempty subset of E. 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 $a \in \partial A$. Otherwise, there would be some open ball $B(a, \epsilon)$ of center a contained in A and so there would be points of A (in $B(a, \epsilon)$) in both half-spaces determined by H, contradicting the fact that H is a supporting hyperplane of A at a. Furthermore, $H \cap \overset{\circ}{A} = \emptyset$.

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! (See Figure 3.4).

Next, we need to define various types of boundary points of closed convex sets.

Definition 3.6. Let E 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

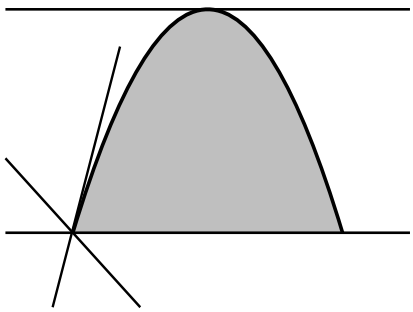


Figure 3.4: Examples of supporting hyperplanes

$a \in \partial A$ 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. See Figure 3.5

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 $0, 1, \ldots, d-1$. This phenomena is illustrated in Figure 3.5. The following proposition is shown in Berger [8] (Proposition 11.6.2):

Proposition 3.6. *The set of vertices of a closed and convex subset is countable.*

Another important concept is that of an extremal point.

Definition 3.7. Let E 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 the interior of any closed nontrivial line segment $[x, y] \subseteq A$ ($x \neq y$, $a \neq x$ and $a \neq y$).

Observe that a vertex is extremal, but the converse is false. For example, in Figure 3.6, all the points on the arc of parabola, including v_1 and v_2, are extreme points. However, only v_1 and v_2 are vertices. Also, if $\dim E \geq 3$, the set of extremal points of a compact convex may not be closed. See Berger [8], Chapter 11, Figure 11.6.5.3, which we reproduce in Figure 3.7.

Actually, it is not at all obvious that a nonempty compact convex set 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. We need the following lemma:

Lemma 3.7. *Let E be an affine space of dimension n, and let A be a nonempty compact and convex set. Then, $A = \mathrm{conv}(\partial A)$, i.e., A is equal to the convex hull of its boundary.*

Proof. Pick any a in A, and consider any line D through a. Then, $D \cap A$ is closed and convex. However, since A is compact, it follows that $D \cap A$ is a closed interval $[u, v]$ containing a, and $u, v \in \partial A$. Therefore, $a \in \mathrm{conv}(\partial A)$, as desired. □

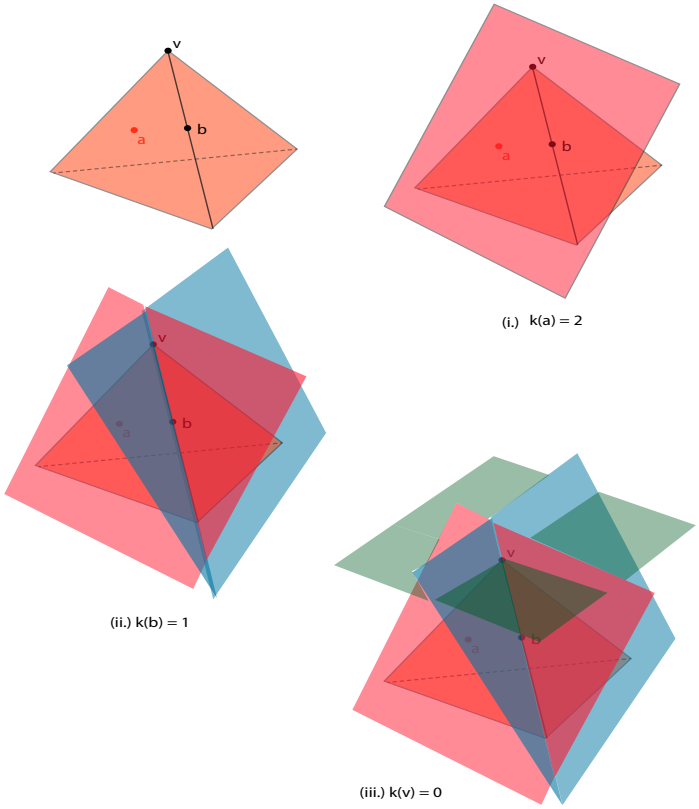


Figure 3.5: The various types of boundary points for a solid tetrahedron. If the point is in the interior of a triangular face, it has order 2. If the point is in the interior of an edge, it has order 1. If the point is a vertex, it has order 0.

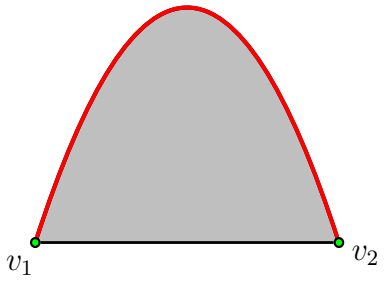


Figure 3.6: Examples of vertices and extreme points

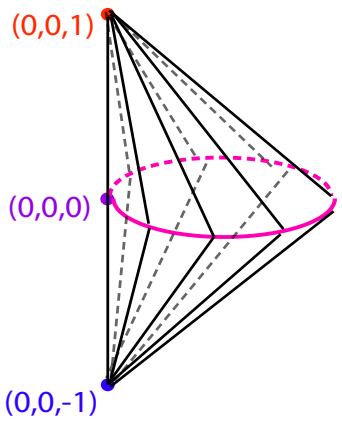


Figure 3.7: Let A the convex set formed by taking a planar unit circle through the origin and forming the double cone with apex $(0, 0, 1)$ and $(0, 0, -1)$. The extremal points of A are the points on pink circular boundary minus the origin.

The following important theorem shows that only extremal points matter as far as determining a compact and convex subset from its boundary. The proof of Theorem 3.8 makes use of a proposition due to Minkowski (Proposition 4.19) which will be proved in Section 4.2.

Theorem 3.8. *(Krein and Milman, 1940) Let E 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.*

Proof. Denote the set of extremal points of A by $\mathrm{Extrem}(A)$. We proceed by induction on $d = \dim E$. When $d = 1$, the convex and compact subset A must be a closed interval $[u, v]$, or a single point. In either cases, the theorem holds trivially. Now, assume $d \geq 2$, and assume that the theorem holds for $d - 1$. It is easily verified that

$$\mathrm{Extrem}(A \cap H) = (\mathrm{Extrem}(A)) \cap H,$$

for every supporting hyperplane H of A (such hyperplanes exist, by Minkowski's proposition (Proposition 4.19)). Observe that Lemma 3.7 implies that if we can prove that

$$\partial A \subseteq \mathrm{conv}(\mathrm{Extrem}(A)),$$

then, since $A = \mathrm{conv}(\partial A)$, we will have established that

$$A = \mathrm{conv}(\mathrm{Extrem}(A)).$$

Let $a \in \partial A$, and let H be a supporting hyperplane of A at a (which exists, by Minkowski's proposition). Now, A and H are convex so $A \cap H$ is convex; H is closed and A is compact,

so $H \cap A$ is a closed subset of a compact subset, A, and thus, $A \cap H$ is also compact. Since $A \cap H$ is a compact and convex subset of H and H has dimension $d-1$, by the induction hypothesis, we have

$$A \cap H = \text{conv}(\text{Extrem}(A \cap H)).$$

However,

$$\begin{aligned} \text{conv}(\text{Extrem}(A \cap H)) &= \text{conv}((\text{Extrem}(A)) \cap H) \\ &= \text{conv}(\text{Extrem}(A)) \cap H \subseteq \text{conv}(\text{Extrem}(A)), \end{aligned}$$

and so, $a \in A \cap H \subseteq \text{conv}(\text{Extrem}(A))$. Therefore, we proved that

$$\partial A \subseteq \text{conv}(\text{Extrem}(A)),$$

from which we deduce that $A = \text{conv}(\text{Extrem}(A))$, as explained earlier. □

Remark: Observe that Krein and Milman's theorem implies that any nonempty 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 applies to infinite dimensional affine spaces, provided that they are locally convex, see Valentine [65], Chapter 11, Bourbaki [13], Chapter II, Barvinok [4], Chapter 3, or Lax [40], Chapter 13.

An important consequence of Krein and Millman's theorem is that every convex function on a convex and compact set achieves its maximum at some extremal point.

Definition 3.8. Let A be a nonempty convex subset of $\mathbb{A}^n$. A function $f\colon A \to \mathbb{R}$ is *convex* if

$$f((1-\lambda)a + \lambda b) \leq (1-\lambda)f(a) + \lambda f(b)$$

for all $a, b \in A$ and for all $\lambda \in [0,1]$. The function $f\colon A \to \mathbb{R}$ is *strictly convex* if

$$f((1-\lambda)a + \lambda b) < (1-\lambda)f(a) + \lambda f(b)$$

for all $a, b \in A$ with $a \neq b$ and for all λ with $0 < \lambda < 1$. A function $f\colon A \to \mathbb{R}$ is *concave* (resp. *strictly concave*) iff $-f$ is convex (resp. $-f$ is strictly convex). See Figure 3.8.

If f is convex, a simple induction shows that

$$f\left(\sum_{i \in I} \lambda_i a_i\right) \leq \sum_{i \in I} \lambda_i f(a_i)$$

for every finite convex combination in A, *i.e.*, for any finite family $(a_i)_{i \in I}$ of points in A and any family $(\lambda_i)_{i \in I}$ with $\sum_{i \in I} \lambda_i = 1$ and $\lambda_i \geq 0$ for all $i \in I$.

Proposition 3.9. *Let A be a nonempty convex and compact subset of $\mathbb{A}^n$ and let $f\colon A \to \mathbb{R}$ be any function. If f is convex and continuous, then f achieves its maximum at some extreme point of A.*

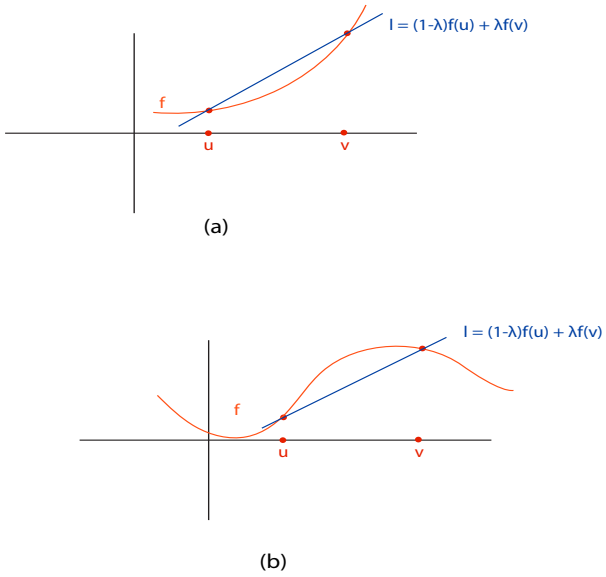


Figure 3.8: Figures (a) and (b) are the graphs of real valued functions. Figure (a) is the graph of convex function since the blue line lies above the graph of f. Figure (b) shows the graph of a function which is not convex.

Proof. Since A is compact and f is continuous, $f(A)$ is a closed interval, $[m, M]$, in $\mathbb{R}$ and so f achieves its minimum m and its maximum M. Say $f(c) = M$, for some $c \in A$. By Krein and Millman's theorem, c is some convex combination of exteme points of A,

$$c = \sum_{i=1}^{k} \lambda_i a_i,$$

with $\sum_{i=1}^{k} \lambda_i = 1$, $\lambda_i \geq 0$ and each a_i an extreme point in A. But then, as f is convex,

$$M = f(c) = f\left(\sum_{i=1}^{k} \lambda_i a_i\right) \leq \sum_{i=1}^{k} \lambda_i f(a_i)$$

and if we let

$$f(a_{i_0}) = \max_{1 \leq i \leq k} \{f(a_i)\}$$

for some i_0 such that $1 \leq i_0 \leq k$, then we get

$$M = f(c) \leq \sum_{i=1}^{k} \lambda_i f(a_i) \leq \left(\sum_{i=1}^{k} \lambda_i\right) f(a_{i_0}) = f(a_{i_0}),$$

as $\sum_{i=1}^{k} \lambda_i = 1$. Since M is the maximum value of the function f over A, we have $f(a_{i_0}) \leq M$ and so,

$$M = f(a_{i_0})$$

and f achieves its maximum at the extreme point a_{i_0}, as claimed. □

Proposition 3.9 plays an important role in convex optimization: It guarantees that the maximum value of a convex objective function on a compact and convex set is achieved at some extreme point. Thus, it is enough to look for a maximum at some extreme point of the domain.

Proposition 3.9 fails for minimal values of a convex function. For example, the function, $x \mapsto f(x) = x^2$, defined on the compact interval $[-1, 1]$ achieves it minimum at $x = 0$, which is not an extreme point of $[-1, 1]$. However, if f is concave, then f achieves its minimum value at some extreme point of A. In particular, if f is affine, it achieves its minimum and its maximum at some extreme points of A.

We conclude this chapter with three other classics of convex geometry.

3.5 Radon's, Tverberg's, Helly's, Theorems and Centerpoints

We begin with *Radon's theorem*.

Theorem 3.10. *(Radon, 1921) 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.*

Proof. Pick some origin O in E. Write $X = (x_i)_{i \in L}$ for some index set L (we can let $L = X$). Since by assumption $|X| \geq m + 2$ where $m = \dim(E)$, X is affinely dependent, and by Lemma 2.6.5 from Gallier [30], there is a family $(\mu_k)_{k \in L}$ (of finite support) of scalars, not all null, such that

$$\sum_{k \in L} \mu_k = 0 \quad \text{and} \quad \sum_{k \in L} \mu_k \mathbf{Ox_k} = 0.$$

Since $\sum_{k \in L} \mu_k = 0$, the μ_k are not all null, and $(\mu_k)_{k \in L}$ has finite support, the sets

$$I = \{i \in L \mid \mu_i > 0\} \quad \text{and} \quad J = \{j \in L \mid \mu_j < 0\}$$

are nonempty, finite, and obviously disjoint. Let

$$X_1 = \{x_i \in X \mid \mu_i > 0\} \quad \text{and} \quad X_2 = \{x_i \in X \mid \mu_i \leq 0\}.$$

Again, since the μ_k are not all null and $\sum_{k \in L} \mu_k = 0$, the sets X_1 and X_2 are nonempty, and obviously

$$X_1 \cap X_2 = \emptyset \quad \text{and} \quad X_1 \cup X_2 = X.$$

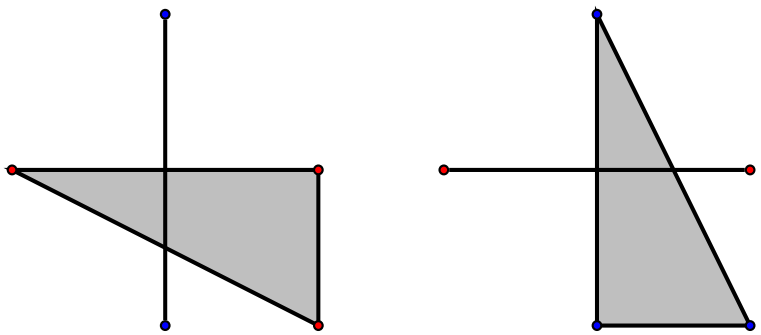


Figure 3.9: Examples of Radon Partitions

Furthermore, the definition of I and J implies that $(x_i)_{i\in I} \subseteq X_1$ and $(x_j)_{j\in J} \subseteq X_2$. It remains to prove that $\text{conv}(X_1) \cap \text{conv}(X_2) \neq \emptyset$. The definition of I and J implies that

$$\sum_{k\in L} \mu_k \mathbf{Ox_k} = 0$$

can be written as

$$\sum_{i\in I} \mu_i \mathbf{Ox_i} + \sum_{j\in J} \mu_j \mathbf{Ox_j} = 0,$$

that is, as

$$\sum_{i\in I} \mu_i \mathbf{Ox_i} = \sum_{j\in J} -\mu_j \mathbf{Ox_j},$$

where

$$\sum_{i\in I} \mu_i = \sum_{j\in J} -\mu_j = \mu,$$

with $\mu > 0$. Thus, we have

$$\sum_{i\in I} \frac{\mu_i}{\mu} \mathbf{Ox_i} = \sum_{j\in J} -\frac{\mu_j}{\mu} \mathbf{Ox_j},$$

with

$$\sum_{i\in I} \frac{\mu_i}{\mu} = \sum_{j\in J} -\frac{\mu_j}{\mu} = 1,$$

proving that $\sum_{i\in I}(\mu_i/\mu)x_i \in \text{conv}(X_1)$ and $\sum_{j\in J} -(\mu_j/\mu)x_j \in \text{conv}(X_2)$ are identical, and thus that $\text{conv}(X_1) \cap \text{conv}(X_2) \neq \emptyset$. □

A partition, (X_1, X_2), of X satisfying the conditions of Theorem 3.10 is sometimes called a *Radon partition* of X and any point in $\text{conv}(X_1) \cap \text{conv}(X_2)$ is called a *Radon point* of X. Figure 3.9 shows two Radon partitions of five points in the plane.

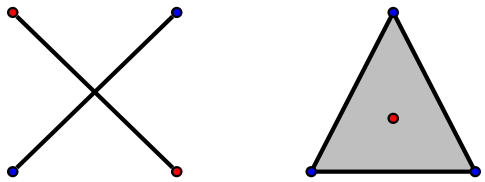


Figure 3.10: The Radon Partitions of four points (in $\mathbb{A}^2$)

It can be shown that a finite set, $X \subseteq E$, has a unique Radon partition iff it has $m+2$ elements and any $m+1$ points of X are affinely independent. For example, there are exactly two possible cases in the plane as shown in Figure 3.10.

There is also a version of Radon's theorem for the class of cones with an apex. Say that a convex cone, $C \subseteq E$, has an *apex* (or is a *pointed cone*) iff there is some hyperplane, H, such that $C \subseteq H_+$ and $H \cap C = \{0\}$. For example, the cone obtained as the intersection of two half spaces in $\mathbb{R}^3$ is not pointed since it is a wedge with a line as part of its boundary. Here is the version of Radon's theorem for convex cones:

Theorem 3.11. *Given any vector space E of dimension m, for every subset X of E, if* $\text{cone}(X)$ *is a pointed cone such that X has at least $m+1$ nonzero vectors, then there is a partition of X into two nonempty disjoint subsets, X_1 and X_2, such that the cones,* $\text{cone}(X_1)$ *and* $\text{cone}(X_2)$*, have a nonempty intersection not reduced to* $\{0\}$*.*

The proof of Theorem 3.11 is left as an exercise.

There is a beautiful generalization of Radon's theorem known as *Tverberg's Theorem*.

Theorem 3.12. *(Tverberg's Theorem, 1966) Let E be any affine space of dimension m. For any natural number, $r \geq 2$, for every subset, X, of E, if X has at least $(m+1)(r-1)+1$ points, then there is a partition, $(X_1, \ldots, X_r)$, of X into r nonempty pairwise disjoint subsets so that* $\bigcap_{i=1}^{r} \text{conv}(X_i) \neq \emptyset$*.*

A partition as in Theorem 3.12 is called a *Tverberg partition* and a point in $\bigcap_{i=1}^{r} \text{conv}(X_i)$ is called a *Tverberg point*. Theorem 3.12 was conjectured by Birch and proved by Tverberg in 1966. Tverberg's original proof was technically quite complicated. Tverberg then gave a simpler proof in 1981 and other simpler proofs were later given, notably by Sarkaria (1992) and Onn (1997), using the Colorful Carathéodory theorem. A proof along those lines can be found in Matousek [41], Chapter 8, Section 8.3. A *colored Tverberg theorem* and more can also be found in Matousek [41] (Section 8.3).

Next, we prove a version of *Helly's theorem*.

Theorem 3.13. *(Helly, 1913) 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.*

Proof. The proof is by induction on $n \geq m+1$ and uses Radon's theorem in the induction step. For $n = m+1$, the assumption of the theorem is that the intersection of any family of $m+1$ of the K_i's is nonempty, and the theorem holds trivially. Next, let $L = \{1, 2, \ldots, n+1\}$, where $n+1 \geq m+2$. By the induction hypothesis, $C_i = \bigcap_{j \in (L-\{i\})} K_j$ is nonempty for every $i \in L$.

We claim that $C_i \cap C_j \neq \emptyset$ for some $i \neq j$. If so, as $C_i \cap C_j = \bigcap_{k=1}^{n+1} K_k$, we are done. So, let us assume that the C_i's are pairwise disjoint. Then, we can pick a set $X = \{a_1, \ldots, a_{n+1}\}$ such that $a_i \in C_i$, for every $i \in L$. By Radon's Theorem, there are two nonempty disjoint sets $X_1, X_2 \subseteq X$ such that $X = X_1 \cup X_2$ and $\mathrm{conv}(X_1) \cap \mathrm{conv}(X_2) \neq \emptyset$. However, $X_1 \subseteq K_j$ for every j with $a_j \notin X_1$. This is because $a_j \notin K_j$ for every j, and so, we get

$$X_1 \subseteq \bigcap_{a_j \notin X_1} K_j.$$

Symetrically, we also have

$$X_2 \subseteq \bigcap_{a_j \notin X_2} K_j.$$

Since the K_j's are convex and

$$\left(\bigcap_{a_j \notin X_1} K_j \right) \cap \left(\bigcap_{a_j \notin X_2} K_j \right) = \bigcap_{i=1}^{n+1} K_i,$$

it follows that $\mathrm{conv}(X_1) \cap \mathrm{conv}(X_2) \subseteq \bigcap_{i=1}^{n+1} K_i$, so that $\bigcap_{i=1}^{n+1} K_i$ is nonempty, contradicting the fact that $C_i \cap C_j = \emptyset$ for all $i \neq j$. □

A more general version of Helly's theorem is proved in Berger [8].

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

To prove this fact, pick a coordinate system in which the y-axis is parallel to the common direction of the parallel lines, and for every line segment S, let

$$CS = \{(\alpha, \beta) \in \mathbb{R}^2 \mid \text{the line } y = \alpha x + \beta \text{ meets } S\}.$$

It is not hard to see that CS is convex. Then, by hypothesis the fact that any three line segments S_i, S_j, S_k meet a line means that $CS_i \cap CS_j \cap CS_k \neq \emptyset$, any Helly's Theorem implies that the family of all the convex sets CS_i has a nonempty intersection, which means that there is a line meeting all the line segments S_i. This situation for four lines is illustrated in Figure 3.11.

We conclude this chapter with a nice application of Helly's Theorem to the existence of centerpoints. Centerpoints generalize the notion of median to higher dimensions. Recall

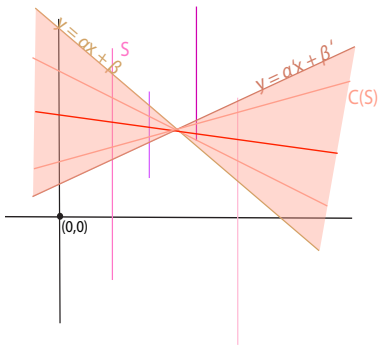


Figure 3.11: The four pink line segments in the affine plane all intersect the horizontal red line.

that if we have a set of n data points, $S = \{a_1, \ldots, a_n\}$, on the real line, a *median* for S is a point, x, such that both intervals $[x, \infty)$ and $(-\infty, x]$ contain at least $n/2$ of the points in S (by $n/2$, we mean the largest integer greater than or equal to $n/2$).

Given any hyperplane, H, recall that the closed half-spaces determined by H are denoted H_+ and H_- and that $H \subseteq H_+$ and $H \subseteq H_-$. We let $\overset{\circ}{H}_+ = H_+ - H$ and $\overset{\circ}{H}_- = H_- - H$ be the *open half-spaces* determined by H.

Definition 3.9. Let $S = \{a_1, \ldots, a_n\}$ be a set of n points in $\mathbb{A}^d$. A point, $c \in \mathbb{A}^d$, is a *centerpoint of* S iff for every hyperplane, H, whenever the closed half-space H_+ (resp. H_-) contains c, then H_+ (resp. H_-) contains at least $\frac{n}{d+1}$ points from S (by $\frac{n}{d+1}$, we mean the largest integer greater than or equal to $\frac{n}{d+1}$, namely the ceiling $\lceil \frac{n}{d+1} \rceil$ of $\frac{n}{d+1}$).

So, for $d = 2$, for each line, D, if the closed half-plane D_+ (resp. D_-) contains c, then D_+ (resp. D_-) contains at least a third of the points from S. For $d = 3$, for each plane, H, if the closed half-space H_+ (resp. H_-) contains c, then H_+ (resp. H_-) contains at least a fourth of the points from S, *etc.* Example 3.12 shows nine points in the plane and one of their centerpoints (in red). This example shows that the bound $\frac{1}{3}$ is tight.

Observe that a point, $c \in \mathbb{A}^d$, is a centerpoint of S iff c belongs to every open half-space, $\overset{\circ}{H}_+$ (resp. $\overset{\circ}{H}_-$) containing at least $\frac{dn}{d+1} + 1$ points from S (again, we mean $\lceil \frac{dn}{d+1} \rceil + 1$).

Indeed, if c is a centerpoint of S and H is any hyperplane such that $\overset{\circ}{H}_+$ (resp. $\overset{\circ}{H}_-$) contains at least $\frac{dn}{d+1} + 1$ points from S, then $\overset{\circ}{H}_+$ (resp. $\overset{\circ}{H}_-$) must contain c as otherwise, the closed half-space, H_- (resp. H_+) would contain c and at most $n - \frac{dn}{d+1} - 1 = \frac{n}{d+1} - 1$ points from S, a contradiction. Conversely, assume that c belongs to every open half-space, $\overset{\circ}{H}_+$ (resp. $\overset{\circ}{H}_-$) containing at least $\frac{dn}{d+1} + 1$ points from S. Then, for any hyperplane, H, if $c \in H_+$ (resp. $c \in H_-$) but H_+ contains at most $\frac{n}{d+1} - 1$ points from S, then the open

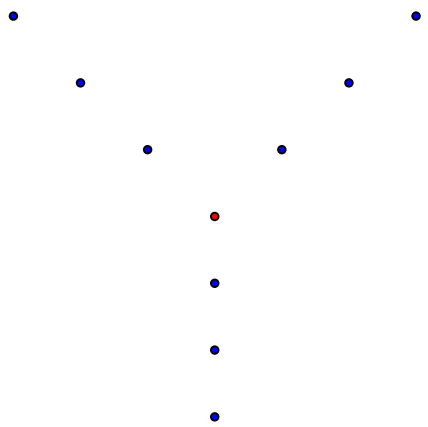


Figure 3.12: Example of a centerpoint

half-space, $\overset{\circ}{H}_-$ (resp. $\overset{\circ}{H}_+$) would contain at least $n - \frac{n}{d+1} + 1 = \frac{dn}{d+1} + 1$ points from S but not c, a contradiction.

We are now ready to prove the existence of centerpoints.

Theorem 3.14. *(Existence of Centerpoints) Every finite set, $S = \{a_1, \ldots, a_n\}$, of n points in $\mathbb{A}^d$ has some centerpoint.*

Proof. We will use the second characterization of centerpoints involving open half-spaces containing at least $\frac{dn}{d+1} + 1$ points.

Consider the family of sets,

$$\begin{aligned}\mathcal{C} &= \left\{\text{conv}(S \cap \overset{\circ}{H}_+) \mid (\exists H)\left(|S \cap \overset{\circ}{H}_+| > \frac{dn}{d+1}\right)\right\} \\ &\quad \cup \left\{\text{conv}(S \cap \overset{\circ}{H}_-) \mid (\exists H)\left(|S \cap \overset{\circ}{H}_-| > \frac{dn}{d+1}\right)\right\},\end{aligned}$$

where H is a hyperplane.

As S is finite, $\mathcal{C}$ consists of a finite number of convex sets, say $\{C_1, \ldots, C_m\}$. If we prove that $\bigcap_{i=1}^m C_i \neq \emptyset$ we are done, because $\bigcap_{i=1}^m C_i$ is the set of centerpoints of S.

First, we prove by induction on k (with $1 \leq k \leq d+1$), that any intersection of k of the C_i's has at least $\frac{(d+1-k)n}{d+1} + k$ elements from S. For $k = 1$, this holds by definition of the C_i's.

Next, consider the intersection of $k+1 \leq d+1$ of the C_i's, say $C_{i_1} \cap \cdots \cap C_{i_k} \cap C_{i_{k+1}}$. Let

$$\begin{aligned}A &= S \cap (C_{i_1} \cap \cdots \cap C_{i_k} \cap C_{i_{k+1}}) \\ B &= S \cap (C_{i_1} \cap \cdots \cap C_{i_k}) \\ C &= S \cap C_{i_{k+1}}.\end{aligned}$$

Note that $A = B \cap C$. By the induction hypothesis, B contains at least $\frac{(d+1-k)n}{d+1} + k$ elements from S. As C contains at least $\frac{dn}{d+1} + 1$ points from S, and as

$$|B \cup C| = |B| + |C| - |B \cap C| = |B| + |C| - |A|$$

and $|B \cup C| \leq n$, we get $n \geq |B| + |C| - |A|$, that is,

$$|A| \geq |B| + |C| - n.$$

It follows that

$$|A| \geq \frac{(d+1-k)n}{d+1} + k + \frac{dn}{d+1} + 1 - n$$

that is,

$$|A| \geq \frac{(d+1-k)n + dn - (d+1)n}{d+1} + k + 1 = \frac{(d+1-(k+1))n}{d+1} + k + 1,$$

establishing the induction hypothesis.

Now, if $m \leq d+1$, the above claim for $k = m$ shows that $\bigcap_{i=1}^{m} C_i \neq \emptyset$ and we are done. If $m \geq d+2$, the above claim for $k = d+1$ shows that any intersection of $d+1$ of the C_i's is nonempty. Consequently, the conditions for applying Helly's Theorem are satisfied and therefore,

$$\bigcap_{i=1}^{m} C_i \neq \emptyset.$$

However, $\bigcap_{i=1}^{m} C_i$ is the set of centerpoints of S and we are done. □

Remark: The above proof actually shows that the set of centerpoints of S is a convex set. In fact, it is a finite intersection of convex hulls of finitely many points, so it is the convex hull of finitely many points, in other words, a polytope. It should also be noted that Theorem 3.14 can be proved easily using Tverberg's theorem (Theorem 3.12). Indeed, for a judicious choice of r, any Tverberg point is a centerpoint!

Jadhav and Mukhopadhyay have given a linear-time algorithm for computing a centerpoint of a finite set of points in the plane. For $d \geq 3$, it appears that the best that can be done (using linear programming) is $O(n^d)$. However, there are good approximation algorithms (Clarkson, Eppstein, Miller, Sturtivant and Teng) and in $\mathbb{E}^3$ there is a near quadratic algorithm (Agarwal, Sharir and Welzl). Recently, Miller and Sheehy (2009) have given an algorithm for finding an approximate centerpoint in sub-exponential time together with a polynomial-checkable proof of the approximation guarantee.

Chapter 4

Two Main Tools: Separation and Polar Duality

4.1 Separation Theorems and Farkas Lemma

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 $\mathbb{A}^d$.

In order to prove the Hahn-Banach theorem, we will need two lemmas. Given any two distinct points $x, y \in X$, we let

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

Our first lemma (Lemma 4.1) is intuitively quite obvious so the reader might be puzzled by the length of its proof. However, after proposing several wrong proofs, we realized that its proof is more subtle than it might appear. The proof below is due to Valentine [65]. See if you can find a shorter (and correct) proof!

Lemma 4.1. *Let S be a nonempty convex set and let $x \in \overset{\circ}{S}$ and $y \in \overline{S}$. Then, we have $]x, y[\subseteq \overset{\circ}{S}$.*

Proof. Let $z \in]x, y[$, that is, $z = (1-\lambda)x + \lambda y$, with $0 < \lambda < 1$. Since $x \in \overset{\circ}{S}$, we can find some open subset, U, contained in S so that $x \in U$. It is easy to check that the central magnification of center z, $H_{z, \frac{\lambda-1}{\lambda}}$, maps x to y. Then, $V = H_{z, \frac{\lambda-1}{\lambda}}(U)$ is an open subset

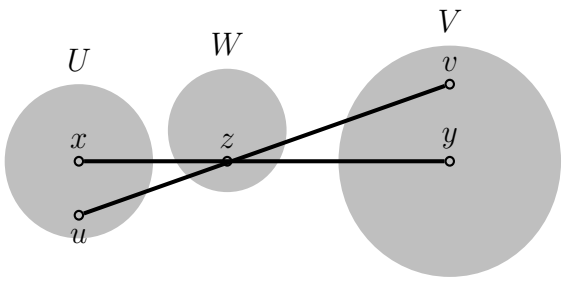


Figure 4.1: Illustration for the proof of Lemma 4.1

containing y and as $y \in \overline{S}$, we have $V \cap S \neq \emptyset$. Let $v \in V \cap S$ be a point of S in this intersection. Now, there is a unique point, $u \in U \subseteq S$, such that $H_{z,\frac{\lambda-1}{\lambda}}(u) = v$ and, as S is convex, we deduce that $z = (1-\lambda)u + \lambda v \in S$. Since U is open, the set

$$W = (1-\lambda)U + \lambda v = \{(1-\lambda)w + \lambda v \mid w \in U\} \subseteq S$$

is also open and $z \in W$, which shows that $z \in \overset{\circ}{S}$. □

Corollary 4.2. *If S is convex, then $\overset{\circ}{S}$ is also convex, and we have $\overset{\circ}{S} = \overline{\overset{\circ}{S}}$. Furthermore, if $\overset{\circ}{S} \neq \emptyset$, then $\overline{S} = \overline{\overset{\circ}{S}}$.*

Beware that if S is a closed set, then the convex hull $\mathrm{conv}(S)$ of S is not necessarily closed!

For example, consider the subset S of $\mathbb{A}^2$ consisting of the points belonging to the right branch of the hyperbola of equation $x^2 - y^2 = 1$, that is,

$$S = \{(x, y) \in \mathbb{R}^2 \mid x^2 - y^2 \geq 1,\ x \geq 0\}.$$

Then S is convex, but the convex hull of the set $S \cup \{(0, 0\}$ is not closed.

However, if S is compact, then $\mathrm{conv}(S)$ is also compact, and thus closed (see Proposition 3.3).

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 (recall that the dimension of a nonempty convex subset is the dimension of its affine hull).

Proposition 4.3. *A nonempty convex set S has a nonempty interior iff* $\dim S = \dim X$.

Proof. Let $d = \dim X$. First, assume that $\overset{\circ}{S} \neq \emptyset$. Then, S contains some open ball of center a_0, and in it, we can find a frame $(a_0, a_1, \ldots, a_d)$ for X. Thus, $\dim S = \dim X$. Conversely, let $(a_0, a_1, \ldots, a_d)$ be a frame of X, with $a_i \in S$, for $i = 0, \ldots, d$. Then, we have

$$\frac{a_0 + \cdots + a_d}{d+1} \in \overset{\circ}{S},$$

and $\overset{\circ}{S}$ is nonempty. □

Proposition 4.3 is false in infinite dimension.

We leave the following property as an exercise:

Proposition 4.4. *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 with vertex x.

Definition 4.1. 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., $H_{x,\lambda}(C) = C$). See Figure 4.2.

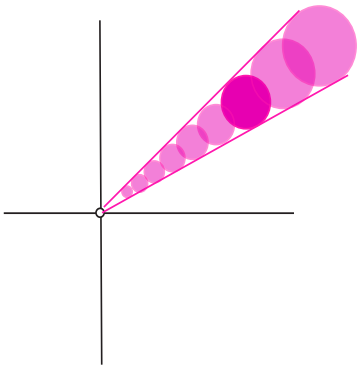


Figure 4.2: For the dark pink disk C, $H_{x,\lambda}(C)$ is the triangular section, excluding O, between the two pink lines.

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 with vertex x.

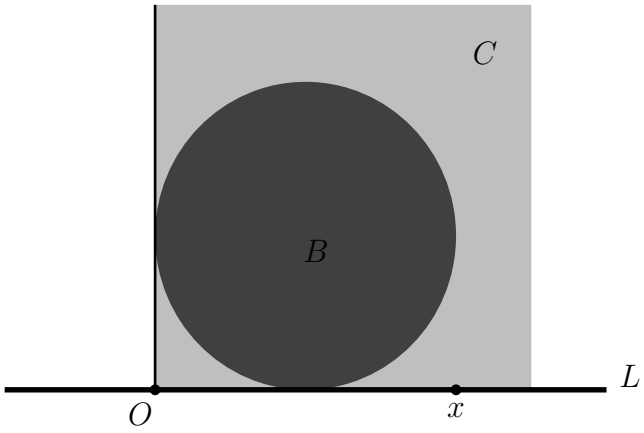


Figure 4.3: Hahn-Banach Theorem in the plane (Lemma 4.5)

Lemma 4.5. *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 $L \cap B = \emptyset$.*

Proof. Define the convex cone $C = \text{cone}_O(B)$. As B is open, it is easy to check that each $H_{O,\lambda}(B)$ is open and since C is the union of the $H_{O,\lambda}(B)$ (for $\lambda > 0$), which are open, C itself is open. Also, $O \notin C$. We claim that at least one point, x, of the boundary, ∂C, of C, is distinct from O. Otherwise, $\partial C = \{O\}$ and we claim that $C = \mathbb{A}^2 - \{O\}$, which is not convex, a contradiction. Indeed, as C is convex it is connected, $\mathbb{A}^2 - \{O\}$ itself is connected and $C \subseteq \mathbb{A}^2 - \{O\}$. If $C \neq \mathbb{A}^2 - \{O\}$, pick some point $a \neq O$ in $\mathbb{A}^2 - C$ and some point $c \in C$. Now, a basic property of connectivity asserts that every continuous path from a (in the exterior of C) to c (in the interior of C) must intersect the boundary of C, namely, $\{O\}$. However, there are plenty of paths from a to c that avoid O, a contradiction. Therefore, $C = \mathbb{A}^2 - \{O\}$.

Since C is open and $x \in \partial C$, we have $x \notin C$. Furthermore, we claim that $y = 2O - x$ (the symmetric of x w.r.t. O) does not belong to C either. Otherwise, we would have $y \in \overset{\circ}{C} = C$ and $x \in \overline{C}$, and by Lemma 4.1, we would get $O \in C$, a contradiction. Therefore, the line through O and x misses C entirely (since C is a cone), and thus, $B \subseteq C$. □

Finally, we come to the Hahn-Banach theorem.

Theorem 4.6. *(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.*

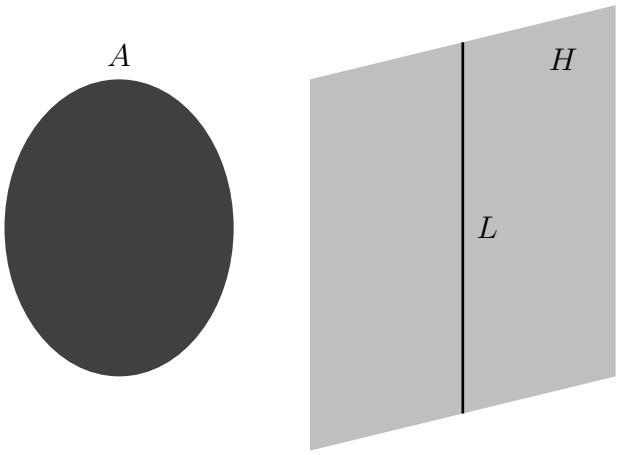


Figure 4.4: Hahn-Banach Theorem, geometric form (Theorem 4.6)

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 $\dim X = 2$. Let V be an affine subspace of X of maximal dimension containing L and so that $V \cap A = \emptyset$. Pick an origin $O \in L$ in X, and consider the vector space X_O. We would like to prove that V is a hyperplane, i.e., $\dim V = \dim X - 1$. We proceed by contradiction. Thus, assume that $\dim V \leq \dim X - 2$. In this case, the quotient space X/V has dimension at least 2. We also know that X/V is isomorphic to the orthogonal complement, $V^\perp$, of V so we may identify X/V and $V^\perp$. The (orthogonal) projection map, $\pi\colon X \to V^\perp$, is linear, continuous, and we can show that π maps the open subset A to an open subset $\pi(A)$, which is also convex (one way to prove that $\pi(A)$ is open is to observe that for any point, $a \in A$, a small open ball of center a contained in A is projected by π to an open ball contained in $\pi(A)$ and as π is surjective, $\pi(A)$ is open). Furthermore, $O \notin \pi(A)$. Since $V^\perp$ has dimension at least 2, there is some plane P (a subspace of dimension 2) intersecting $\pi(A)$, and thus, we obtain a nonempty open and convex subset $B = \pi(A) \cap P$ in the plane $P \cong \mathbb{A}^2$. So, we can apply Lemma 4.5 to B and the point $O = 0$ in $P \cong \mathbb{A}^2$ to find a line, l, (in P) through O with $l \cap B = \emptyset$. But then, $l \cap \pi(A) = \emptyset$ and $W = \pi^{-1}(l)$ is an affine subspace such that $W \cap A = \emptyset$ and W properly contains V, contradicting the maximality of V. See Figure 4.5. □

Remark: The geometric form of the Hahn-Banach theorem also holds when the dimension of X is infinite but a slightly more sophisticated proof is required. Actually, all that is needed is to prove that a maximal affine subspace containing L and disjoint from A exists. This can be done using Zorn's lemma. For other proofs, see Bourbaki [13], Chapter 2, Valentine [65], Chapter 2, Barvinok [4], Chapter 2, or Lax [40], Chapter 3.

Theorem 4.6 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 closed

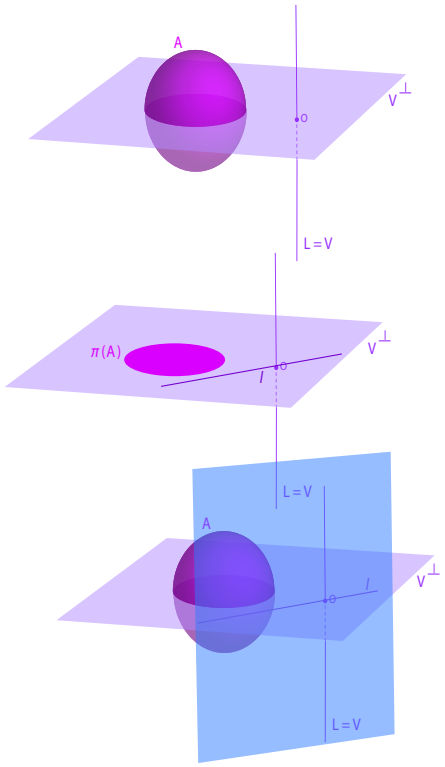


Figure 4.5: An illustration of the proof of 4.6. Let $X = \mathbb{A}^3$, A be the open spherical ball, and L the vertical purple line. The blue hyperplane, which strictly separates A from L is construction using $V^\perp$ and l.

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 as shown by the following counter-example.

In $\mathbb{E}^3$, consider the closed convex set (cone) A defined by the inequalities

$$x \geq 0, \quad y \geq 0, \quad z \geq 0, \quad z^2 \leq xy,$$

and let D be the line given by $x = 0, z = 1$. Then $D \cap A = \emptyset$, both A and D are convex and closed, yet every plane containing D meets A.

Theorem 4.6 has many important corollaries. For example, 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 (recall the definition of a separating hyperplane given in Definition 3.4). We begin with the following version of the Hahn-Banach theorem:

Theorem 4.7. *(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.*

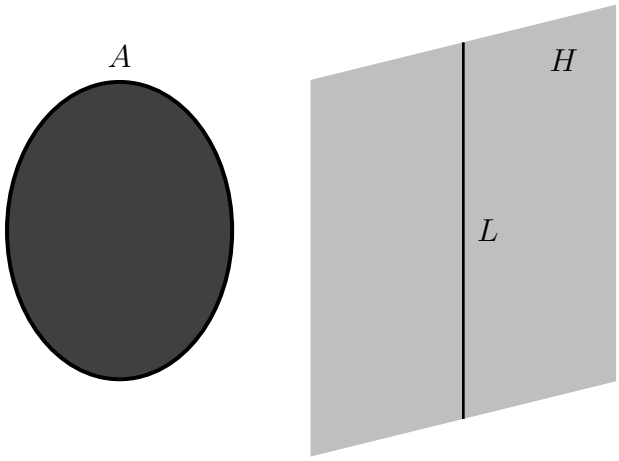


Figure 4.6: Hahn-Banach Theorem, second version (Theorem 4.7)

Proof. Since A is convex, by Corollary 4.2, $\overset{\circ}{A}$ is also convex. By hypothesis, $\overset{\circ}{A}$ is nonempty. So, we can apply Theorem 4.6 to the nonempty open and convex $\overset{\circ}{A}$ and to the affine subspace L. We get a hyperplane H containing L such that $\overset{\circ}{A} \cap H = \emptyset$. However, $A \subseteq \overline{A} = \overline{\overset{\circ}{A}}$ and $\overline{\overset{\circ}{A}}$ is contained in the closed half space (H_+ or H_-) containing $\overset{\circ}{A}$, so H separates A and L. □

Corollary 4.8. *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.*

Proof. Pick some origin O and consider the vector space X_O. Define $C = A - B$ (a special case of the Minkowski sum) as follows:

$$A - B = \{a - b \mid a \in A,\ b \in B\} = \bigcup_{b \in B} (A - b).$$

It is easily verified that $C = A - B$ is convex and has nonempty interior (as a union of subsets having a nonempty interior). Furthermore $O \notin C$, since $A \cap B = \emptyset$.[1] (Note that the definition depends on the choice of O, but this has no effect on the proof.) Since $\overset{\circ}{C}$ is nonempty, we can apply Theorem 4.7 to C and to the affine subspace $\{O\}$ and we get a hyperplane, H,

[1]Readers who prefer a purely affine argument may define $C = A - B$ as the *affine* subset

$$A - B = \{O + a - b \mid a \in A,\ b \in B\}.$$

Again, $O \notin C$ and C is convex. We can pick the affine form, f, defining a separating hyperplane, H, of C and $\{O\}$, so that $f(O + a - b) \leq f(O)$, for all $a \in A$ and all $b \in B$, *i.e.*, $f(a) \leq f(b)$.

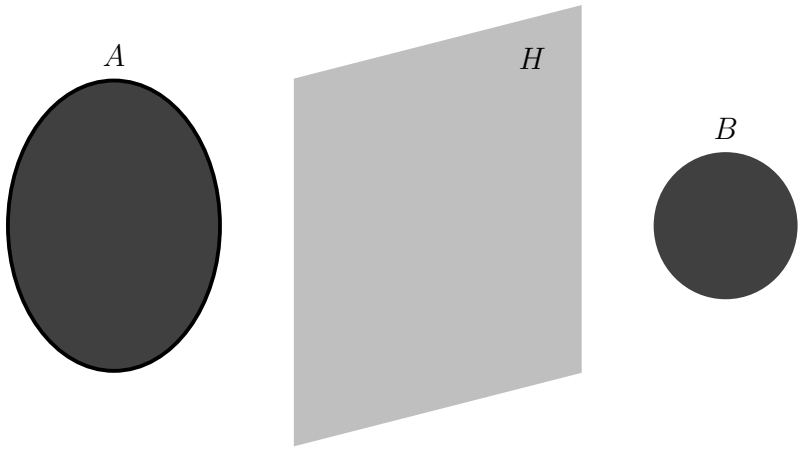


Figure 4.7: Separation Theorem, version 1 (Corollary 4.8)

separating C and $\{O\}$. Let f be any linear form defining the hyperplane H. We may assume that $f(a-b) \leq 0$, for all $a \in A$ and all $b \in B$, *i.e.*, $f(a) \leq f(b)$. Consequently, if we let $\alpha = \sup\{f(a) \mid a \in A\}$ (which makes sense, since the set $\{f(a) \mid a \in A\}$ is bounded), we have $f(a) \leq \alpha$ for all $a \in A$ and $f(b) \geq \alpha$ for all $b \in B$, which shows that the affine hyperplane defined by $f - \alpha$ separates A and B. □

Remark: Theorem 4.7 and Corollary 4.8 also hold in the infinite dimensional case, see Lax [40], Chapter 3, or Barvinok, Chapter 3.

Since a hyperplane, H, separating A and B as in Corollary 4.8 is the boundary of each of the two half–spaces that it determines, we also obtain the following corollary:

Corollary 4.9. *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 4.9 *fails* for *closed* convex sets.

However, Corollary 4.9 holds if we also assume that A (or B) is compact, as shown in Corollary 4.10.

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

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

The proof uses the fact that the distance function is continuous and that a continuous function attains its minimum on a compact set, and is left as an exercise.

Corollary 4.10. *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.*

Proof sketch. First, we pick an origin O and we give $X_O \cong \mathbb{A}^n$ a Euclidean structure. Let d denote the associated distance. Given any subsets A of X, let

$$A + B(O, \epsilon) = \{x \in X \mid d(x, A) < \epsilon\},$$

where $B(a, \epsilon)$ denotes the open ball, $B(a, \epsilon) = \{x \in X \mid d(a, x) < \epsilon\}$, of center a and radius $\epsilon > 0$. Note that

$$A + B(O, \epsilon) = \bigcup_{a \in A} B(a, \epsilon),$$

which shows that $A + B(O, \epsilon)$ is open; furthermore it is easy to see that if A is convex, then $A + B(O, \epsilon)$ is also convex. Now, the function $a \mapsto d(a, B)$ (where $a \in A$) is continuous and since A is compact, it achieves its minimum, $d(A, B) = \min_{a \in A} d(a, B)$, at some point, a, of A. Say, $d(A, B) = \delta$. Since B is closed, there is some $b \in B$ so that $d(A, B) = d(a, B) = d(a, b)$ and since $A \cap B = \emptyset$, we must have $\delta > 0$. Thus, if we pick $\epsilon < \delta/2$, we see that

$$(A + B(O, \epsilon)) \cap (B + B(O, \epsilon)) = \emptyset.$$

Now, $A + B(O, \epsilon)$ and $B + B(O, \epsilon)$ are open, convex and disjoint and we conclude by applying Corollary 4.9. □

Finally, we have the separation theorem announced earlier for arbitrary nonempty convex subsets.

Theorem 4.11. *(Separation of disjoint convex sets) Given an affine space, X, let A and B be two nonempty disjoint convex subsets. Then, there is a hyperplane separating A and B.*

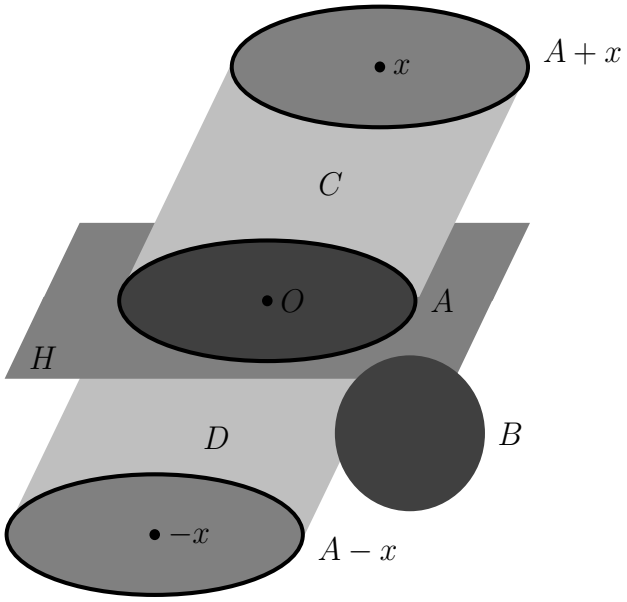


Figure 4.8: Separation Theorem, final version (Theorem 4.11)

Proof. The proof is by descending induction on $\dim A$. If $\dim A = \dim X$, we know from Proposition 4.3 that A has nonempty interior and we conclude using Corollary 4.8. Next, asssume that the induction hypothesis holds if $\dim A \geq n$ and assume $\dim A = n - 1$. Pick an origin $O \in A$ and let H be a hyperplane containing A. Pick $x \in X$ outside H and define $C = \text{conv}(A \cup \{A + x\})$ where $A + x = \{a + x \mid a \in A\}$ and $D = \text{conv}(A \cup \{A - x\})$ where $A - x = \{a - x \mid a \in A\}$. Note that $C \cup D$ is convex. If $B \cap C \neq \emptyset$ and $B \cap D \neq \emptyset$, then the convexity of B and $C \cup D$ implies that $A \cap B \neq \emptyset$, a contradiction. Without loss of generality, assume that $B \cap C = \emptyset$. Since x is outside H, we have $\dim C = n$ and by the induction hypothesis, there is a hyperplane, H_1 separating C and B. As $A \subseteq C$, we see that H_1 also separates A and B. □

Remarks:

(1) The reader should compare this proof (from Valentine [65], Chapter II) with Berger's proof using compactness of the projective space $\mathbb{P}^d$, see Berger [8] (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 [65] (Theorem 2.4):

Lemma 4.12. *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 4.8 and Theorem 4.11. Yet another approach is followed in Barvinok [4].

(3) How can some of the above results be generalized to infinite dimensional affine spaces, especially Theorem 4.6 and Corollary 4.8? 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 [65].

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 $\operatorname{lina} A$ be the set of all points linearly accessible from A and $\operatorname{lin} A = A \cup \operatorname{lina} A$.

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 $\operatorname{core} A$.

It is not difficult to prove that $\operatorname{lin} A \subseteq \overline{A}$ and $\overset{\circ}{A} \subseteq \operatorname{core} A$. If A has nonempty interior, then $\operatorname{lin} A = \overline{A}$ and $\overset{\circ}{A} = \operatorname{core} A$. Also, if A is convex, then $\operatorname{core} A$ and $\operatorname{lin} A$ are convex. Then, Lemma 4.12 still holds (where X is not necessarily finite dimensional) if we redefine V as $V = \operatorname{lin} A \cap \operatorname{lin} B$ and allow the possibility that V could be X itself. Corollary 4.8 also holds in the general case if we assume that $\operatorname{core} A$ is nonempty. For details, see Valentine [65], Chapter I and II.

(4) Yet another approach is to define the notion of an algebraically open convex set, as in Barvinok [4]. 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 4.8 also holds provided A is algebraically open. For details, see Barvinok [4], Chapter 2 and 3. We do not know how the notion "algebraically open" relates to the concept of core.

(5) Theorems 4.6, 4.7 and Corollary 4.8 are proved in Lax [40] 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.

A "cute" application of Corollary 4.10 is one of the many versions of "Farkas Lemma" (1893-1894, 1902), a basic result in the theory of linear programming. For any vector, $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, and any real, $\alpha \in \mathbb{R}$, write $x \geq \alpha$ iff $x_i \geq \alpha$, for $i = 1, \ldots, n$.

The proof of Farkas Lemma Version I (Proposition 4.14) relies on the fact that a polyhedral cone $\operatorname{cone}(a_1, \ldots, a_m)$ is closed. Although it seems obvious that a polyhedral cone should be closed, a rigorous proof is not entirely trivial.

Indeed, the fact that a polyhedral cone is closed relies crucially on the fact that C is spanned by a finite number of vectors, because the cone generated by an infinite set may not be closed. For example, consider the closed disk $D \subseteq \mathbb{R}^2$ of center $(0, 1)$ and radius 1, which is tangent to the x-axis at the origin. Then the cone(D) consists of the open upper half-plane *plus* the origin $(0, 0)$, but this set is not closed.

Proposition 4.13. *Every polyhedral cone C is closed.*

Proof. This is proved by showing that

1. Every primitive cone is closed.

2. A polyhedral cone C is the union of finitely many primitive cones, where a *primitive cone* is a polyhedral cone spanned by linearly independent vectors.

Assume that $(a_1, \ldots, a_m)$ are linearly independent vectors in $\mathbb{R}^n$, and consider any sequence $(x^{(k)})_{k\geq 0}$

$$x^{(k)} = \sum_{i=1}^{m} \lambda_i^{(k)} a_i$$

of vectors in the primitive cone cone$(\{a_1, \ldots, a_m\})$, which means that $\lambda_j^{(k)} \geq 0$ for $i = 1, \ldots, m$ and all $k \geq 0$. The vectors $x^{(k)}$ belong to the subspace U spanned by $(a_1, \ldots, a_m)$, and U is closed. Assume that the sequence $(x^{(k)})_{k\geq 0}$ converges to a limit $x \in \mathbb{R}^n$. Since U is closed and $x^{(k)} \in U$ for all $k \geq 0$, we have $x \in U$. If we write $x = x_1 a_1 + \cdots + x_m a_m$, we would like to prove that $x_i \geq 0$ for $i = 1, \ldots, m$. The sequence the $(x^{(k)})_{k\geq 0}$ converges to x iff

$$\lim_{k\mapsto\infty} \left\|x^{(k)} - x\right\| = 0,$$

iff

$$\lim_{k\mapsto\infty} \Big(\sum_{i=1}^{m} |\lambda_i^{(k)} - x_i|^2\Big)^{1/2} = 0$$

iff

$$\lim_{k\mapsto\infty} \lambda_i^{(k)} = x_i, \quad i = 1, \ldots, m.$$

Since $\lambda_i^{(k)} \geq 0$ for $i = 1, \ldots, m$ and all $k \geq 0$, we have $x_i \geq 0$ for $i = 1, \ldots, m$, so $x \in \text{cone}(\{a_1, \ldots, a_m\})$.

Next, assume that x belongs to the polyhedral cone C. Consider a positive combination

$$x = \lambda_1 a_1 + \cdots + \lambda_k a_k, \tag{$*_1$}$$

for some nonzero $a_1, \ldots, a_k \in C$, with $\lambda_i \geq 0$ and with k *minimal*. Since k is minimal, we must have $\lambda_i > 0$ for $i = 1, \ldots, k$. We claim that $(a_1, \ldots, a_k)$ are linearly independent.

If not, there is some nontrivial linear combination

$$\mu_1 a_1 + \cdots + \mu_k a_k = 0, \tag{$*_2$}$$

and since the a_i are nonzero, $\mu_j \neq 0$ for some at least some j. We may assume that $\mu_j < 0$ for some j (otherwise, we consider the family $(-\mu_i)_{1 \leq i \leq k}$), so let

$$J = \{j \in \{1, \ldots, k\} \mid \mu_j < 0\}.$$

For any $t \in \mathbb{R}$, since $x = \lambda_1 a_1 + \cdots + \lambda_k a_k$, using $(*_2)$ we get

$$x = (\lambda_1 + t\mu_1)a_1 + \cdots + (\lambda_k + t\mu_k)a_k, \tag{$*_3$}$$

and if we pick

$$t = \min_{j \in J}\left(-\frac{\lambda_j}{\mu_j}\right) \geq 0,$$

we have $(\lambda_i + t\mu_i) \geq 0$ for $i = 1, \ldots, k$, but $\lambda_j + t\mu_j = 0$ for some $j \in J$, so $(*_3)$ is an expression of x with less that k nonzero coefficients, contadicting the minimality of k in $(*_1)$. Therefore, $(a_1, \ldots, a_k)$ are linearly independent.

Since a polyhedral cone C is spanned by finitely many vectors, there are finitely many primitive cones (corresponding to linearly independent subfamilies), and since every $x \in C$, belongs to some primitive cone, C is the union of a finite number of primitive cones. Since every primitive cone is closed, as a union of finitely many closed sets, C itself is closed. □

Lemma 4.14. *(Farkas Lemma, Version I) Given any $d \times n$ real matrix A, and any vector $z \in \mathbb{R}^d$, exactly one of the following alternatives occurs:*

(a) The linear system $Ax = z$ has a solution $x = (x_1, \ldots, x_n)$, such that $x \geq 0$ and $x_1 + \cdots + x_n = 1$, or

(b) There is some $c \in \mathbb{R}^d$ and some $\alpha \in \mathbb{R}$ such that $c^\top z < \alpha$ and $c^\top A \geq \alpha$.

Proof. Let $A_1, \ldots, A_n \in \mathbb{R}^d$ be the n points corresponding to the columns of A. Then, either $z \in \mathrm{conv}(\{A_1, \ldots, A_n\})$ or $z \notin \mathrm{conv}(\{A_1, \ldots, A_n\})$. In the first case, we have a convex combination

$$z = x_1 A_1 + \cdots + x_n A_n$$

where $x_i \geq 0$ and $x_1 + \cdots + x_n = 1$, so $x = (x_1, \ldots, x_n)$ is a solution satisfying (a).

In the second case, by Corollary 4.10, there is a hyperplane, H, strictly separating $\{z\}$ and $\mathrm{conv}(\{A_1, \ldots, A_n\})$, which is closed by Proposition 4.13. In fact, observe that $z \notin \mathrm{conv}(\{A_1, \ldots, A_n\})$ iff there is a hyperplane, H, such that $z \in \overset{\circ}{H}_-$ and $A_i \in H_+$, or $z \in \overset{\circ}{H}_+$ and $A_i \in H_-$, for $i = 1, \ldots, n$. As the affine hyperplane, H, is the zero locus of an equation of the form

$$c_1 y_1 + \cdots + c_d y_d = \alpha,$$

either $c^\top z < \alpha$ and $c^\top A_i \geq \alpha$ for $i = 1, \ldots, n$, that is, $c^\top A \geq \alpha$, or $c^\top z > \alpha$ and $c^\top A \leq \alpha$. In the second case, $(-c)^\top z < -\alpha$ and $(-c)^\top A \geq -\alpha$, so (b) is satisfied by either c and α or by $-c$ and $-\alpha$. □

Remark: If we relax the requirements on solutions of $Ax = z$ and only require $x \geq 0$ ($x_1 + \cdots + x_n = 1$ is no longer required) then, in condition (b), we can take $\alpha = 0$. This is another version of Farkas Lemma. In this case, instead of considering the convex hull of $\{A_1, \ldots, A_n\}$ we are considering the convex cone,

$$\text{cone}(A_1, \ldots, A_n) = \{\lambda A_1 + \cdots + \lambda_n A_n \mid \lambda_i \geq 0,\ 1 \leq i \leq n\},$$

that is, we are dropping the condition $\lambda_1 + \cdots + \lambda_n = 1$. For this version of Farkas Lemma we need the following separation lemma:

Proposition 4.15. *Let $C \subseteq \mathbb{E}^d$ be any closed convex cone with vertex O. Then, for every point a not in C, there is a hyperplane H passing through O separating a and C with $a \notin H$.*

Proof. Since C is closed and convex and $\{a\}$ is compact and convex, by Corollary 4.10, there is a hyperplane, H', strictly separating a and C. Let H be the hyperplane through O parallel to H'. Since C and a lie in the two disjoint open half-spaces determined by H', the point a cannot belong to H. Suppose that some point, $b \in C$, lies in the open half-space determined by H and a. Then, the line, L, through O and b intersects H' in some point, c, and as C is a cone, the half line determined by O and b is contained in C. So, $c \in C$ would belong to H', a contradiction. Therefore, C is contained in the closed half-space determined by H that does not contain a, as claimed. □

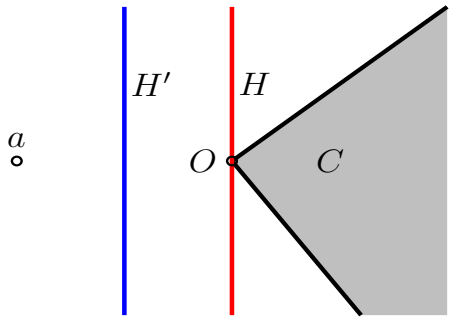


Figure 4.9: Illustration for the proof of Proposition 4.15

Lemma 4.16. *(Farkas Lemma, Version II) Given any $d \times n$ real matrix A and any vector $z \in \mathbb{R}^d$, exactly one of the following alternatives occurs:*

(a) The linear system $Ax = z$ has a solution x such that $x \geq 0$, or

(b) There is some $c \in \mathbb{R}^d$ such that $c^\top z < 0$ and $c^\top A \geq 0$.

Proof. The proof is analogous to the proof of Lemma 4.14 except that it uses Proposition 4.15 instead of Corollary 4.10 and either $z \in \text{cone}(A_1, \ldots, A_n)$ or $z \notin \text{cone}(A_1, \ldots, A_n)$. □

One can show that Farkas II implies Farkas I. Here is another version of Farkas Lemma having to do with a system of inequalities, $Ax \leq z$. Although, this version may seem weaker that Farkas II, it is actually equivalent to it!

Lemma 4.17. *(Farkas Lemma, Version III) Given any $d \times n$ real matrix A and any vector $z \in \mathbb{R}^d$, exactly one of the following alternatives occurs:*

(a) The system of inequalities $Ax \leq z$ has a solution x, or

(b) There is some $c \in \mathbb{R}^d$ such that $c \geq 0$, $c^\top z < 0$ and $c^\top A = 0$.

Proof. We use two tricks from linear programming:

1. We convert the system of inequalities, $Ax \leq z$, into a system of equations by introducing a vector of "slack variables", $\gamma = (\gamma_1, \ldots, \gamma_d)$, where the system of equations is
$$(A, I)\begin{pmatrix} x \\ \gamma \end{pmatrix} = z,$$
with $\gamma \geq 0$.

2. We replace each "unconstrained variable", x_i, by $x_i = X_i - Y_i$, with $X_i, Y_i \geq 0$.

Then, the original system $Ax \leq z$ has a solution, x (unconstrained), iff the system of equations
$$(A, -A, I)\begin{pmatrix} X \\ Y \\ \gamma \end{pmatrix} = z$$
has a solution with $X, Y, \gamma \geq 0$. By Farkas II, this system has no solution iff there exists some $c \in \mathbb{R}^d$ with $c^\top z < 0$ and
$$c^\top (A, -A, I) \geq 0,$$
that is, $c^\top A \geq 0$, $-c^\top A \geq 0$, and $c \geq 0$. However, these four conditions reduce to $c^\top z < 0$, $c^\top A = 0$ and $c \geq 0$. □

These versions of Farkas lemma are statements of the form $(P \vee Q) \wedge \neg(P \wedge Q)$, which is easily seen to be equivalent to $\neg P \equiv Q$, namely, the logical equivalence of $\neg P$ and Q. Therefore, Farkas-type lemmas can be interpreted as criteria for the unsolvablity of various kinds of systems of linear equations or systems of linear inequalities, in the form of a separation property.

For example, Farkas II (Lemma 4.16) says that a system of linear equations, $Ax = z$, does not have any solution, $x \geq 0$, iff there is some $c \in \mathbb{R}^d$ such that $c^\top z < 0$ and $c^\top A \geq 0$. This means that there is a hyperplane, H, of equation $c^\top y = 0$, such that the columns vectors, A_j, forming the matrix A all lie in the positive closed half space, H_+, but z lies in the interior of the other half space, H_-, determined by H. Therefore, z can't be in the cone spanned by the A_j's.

Farkas III says that a system of linear inequalities, $Ax \leq z$, does not have any solution (at all) iff there is some $c \in \mathbb{R}^d$ such that $c \geq 0$, $c^\top z < 0$ and $c^\top A = 0$. This time, there is also a hyperplane of equation $c^\top y = 0$, with $c \geq 0$, such that the columns vectors, A_j, forming the matrix A all lie in H but z lies in the interior of the half space, H_-, determined by H. In the "easy" direction, if there is such a vector c and some x satisfying $Ax \leq z$, since $c \geq 0$, we get $c^\top Ax \leq c^\top z$, but $c^\top Ax = 0$ and $c^\top z < 0$, a contradiction.

What is the crirerion for the insolvability of a system of inequalities $Ax \leq z$ with $x \geq 0$? This problem is equivalent to the insolvability of the set of inequalities

$$\begin{pmatrix} A \\ -I \end{pmatrix} x \leq \begin{pmatrix} z \\ 0 \end{pmatrix}$$

and by Farkas III, this system has no solution iff there is some vector, (c_1, c_2), with $(c_1, c_2) \geq 0$,

$$(c_1^\top, c_2^\top) \begin{pmatrix} A \\ -I \end{pmatrix} = 0 \quad \text{and} \quad (c_1^\top, c_2^\top) \begin{pmatrix} z \\ 0 \end{pmatrix} < 0.$$

The above conditions are equivalent to $c_1 \geq 0$, $c_2 \geq 0$, $c_1^\top A - c_2^\top = 0$ and $c_1^\top z < 0$, which reduce to $c_1 \geq 0$, $c_1^\top A \geq 0$ and $c_1^\top z < 0$.

We can put all these versions together to prove the following version of Farkas lemma:

Lemma 4.18. *(Farkas Lemma, Version IIIb) For any $d \times n$ real matrix A and any vector $z \in \mathbb{R}^d$, the following statements are equivalent:*

(1) The system $Ax = z$ has no solution $x \geq 0$ iff there is some $c \in \mathbb{R}^d$ such that $c^\top A \geq 0$ and $c^\top z < 0$.

(2) The system $Ax \leq z$ has no solution iff there is some $c \in \mathbb{R}^d$ such that $c \geq 0$, $c^\top A = 0$ and $c^\top z < 0$.

(3) The system $Ax \leq z$ has no solution $x \geq 0$ iff there is some $c \in \mathbb{R}^d$ such that $c \geq 0$, $c^\top A \geq 0$ and $c^\top z < 0$.

Proof. We already proved that (1) implies (2) and that (2) implies (3). The proof that (3) implies (1) is left as an easy exercise. □

The reader might wonder what is the criterion for the unsolvability of a system $Ax = z$, without any condition on x. However, since the unsolvability of the system $Ax = b$ is equivalent to the unsolvability of the system

$$\begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} z \\ -z \end{pmatrix},$$

using (2), the above system is unsolvable iff there is some $(c_1, c_2) \geq (0, 0)$ such that

$$(c_1^\top, c_2^\top) \begin{pmatrix} A \\ -A \end{pmatrix} = 0 \quad \text{and} \quad (c_1^\top, c_2^\top) \begin{pmatrix} z \\ -z \end{pmatrix} < 0,$$

and these are equivalent to $c_1^\top A - c_2^\top A = 0$ and $c_1^\top z - c_2^\top z < 0$, namely, $c^\top A = 0$ and $c^\top z < 0$ where $c = c_1 - c_2 \in \mathbb{R}^d$. However, this simply says that c is orthogonal to the columns $A^1, \ldots, A^n$ of A and that z is not orthogonal to c, so z cannot belong to the column space of A, a criterion that we already knew from linear algebra.

As in Matousek and Gartner [42], we can summarize these various criteria in the following table:

	The system $Ax \leq z$	The system $Ax = z$
has no solution $x \geq 0$ iff	$\exists c \in \mathbb{R}^d$, such that $c \geq 0$, $c^\top A \geq 0$ and $c^\top z < 0$	$\exists c \in \mathbb{R}^d$, such that $c^\top A \geq 0$ and $c^\top z < 0$
has no solution $x \in \mathbb{R}^n$ iff	$\exists c \in \mathbb{R}^d$, such that, $c \geq 0$, $c^\top A = 0$ and $c^\top z < 0$	$\exists c \in \mathbb{R}^d$, such that $c^\top A = 0$ and $c^\top z < 0$

Remark: The strong duality theorem in linear programming can be proved using Lemma 4.18(c).

4.2 Supporting Hyperplanes and Minkowski's Proposition

Recall the definition of a supporting hyperplane given in Definition 3.5. We have the following important proposition first proved by Minkowski (1896):

Proposition 4.19. *(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.*

Proof. Let $d = \dim A$. If $d < \dim X$ (i.e., A has empty interior), then A is contained in some affine subspace V of dimension $d < \dim X$, and any hyperplane containing V is a supporting hyperplane for every $a \in A$. Now, assume $d = \dim X$, so that $\overset{\circ}{A} \neq \emptyset$. If $a \in \partial A$, then $\{a\} \cap \overset{\circ}{A} = \emptyset$. By Theorem 4.6, there is a hyperplane H separating $\overset{\circ}{A}$ and $L = \{a\}$. However, by Corollary 4.2, since $\overset{\circ}{A} \neq \emptyset$ and A is closed, we have

$$A = \overline{A} = \overline{\overset{\circ}{A}}.$$

Now, the half–space containing $\overset{\circ}{A}$ is closed, and thus, it contains $\overline{\overset{\circ}{A}} = A$. Therefore, H separates A and $\{a\}$. □

Remark: The assumption that A is closed is convenient but unnecessary. Indeed, the proof of Proposition 4.19 shows that the proposition holds for every boundary point, $a \in \partial A$ (assuming $\partial A \neq \emptyset$).

Beware that Proposition 4.19 is false when the dimension of X is infinite and when $\overset{\circ}{A} = \emptyset$.

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

Proposition 4.20. *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.*

Proof. We leave it as an exercise (see Berger [8], Proposition 11.5.4). □

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

Proof. Let A' be the intersection of all the closed half–spaces containing A. It is immediately checked that A' is closed and convex and that $A \subseteq A'$. Assume that $A' \neq A$, and pick $a \in A' - A$. Then, we can apply Corollary 4.10 to $\{a\}$ and A and we find a hyperplane, H, strictly separating A and $\{a\}$; this shows that A belongs to one of the two half-spaces determined by H, yet a does not belong to the same half-space, contradicting the definition of A'. □

4.3 Polarity and Duality

Let $E = \mathbb{E}^n$ be the Euclidean affine space of dimension n. We will denote the origin $(0, \ldots, 0)$ in $\mathbb{E}^n$ by O. We know that the inner product on $E = \mathbb{E}^n$ induces a duality between E and its dual E^* (for example, see Chapter 6, Section 2 of Gallier [30]), 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

$$H = \{a \in \mathbb{E}^n \mid \mathbf{Oh} \cdot \mathbf{Oa} = 1\}.$$

Indeed, any hyperplane H in $\mathbb{E}^n$ is the null set of some equation of the form

$$\alpha_1 x_1 + \cdots + \alpha_n x_n = \beta,$$

and if $O \notin H$, then $\beta \neq 0$. Thus, any hyperplane H not passing through O is defined by an equation of the form

$$h_1 x_1 + \cdots + h_n x_n = 1,$$

if we set $h_i = \alpha_i/\beta$. So, if we let $h = (h_1, \ldots, h_n)$, we see that

$$H = \{a \in \mathbb{E}^n \mid \mathbf{Oh} \cdot \mathbf{Oa} = 1\},$$

as claimed. Now, assume that

$$H = \{a \in \mathbb{E}^n \mid \mathbf{Oh_1} \cdot \mathbf{Oa} = 1\} = \{a \in \mathbb{E}^n \mid \mathbf{Oh_2} \cdot \mathbf{Oa} = 1\}.$$

The functions $a \mapsto \mathbf{Oh_1} \cdot \mathbf{Oa} - 1$ and $a \mapsto \mathbf{Oh_2} \cdot \mathbf{Oa} - 1$ are two affine forms defining the same hyperplane, so there is a nonzero scalar λ so that

$$\mathbf{Oh_1} \cdot \mathbf{Oa} - 1 = \lambda(\mathbf{Oh_2} \cdot \mathbf{Oa} - 1) \quad \text{for all } a \in \mathbb{E}^n$$

(see Gallier [30], Chapter 2, Section 2.10). In particular, for $a = O$, we find that $\lambda = 1$, and so,

$$\mathbf{Oh_1} \cdot \mathbf{Oa} = \mathbf{Oh_2} \cdot \mathbf{Oa} \quad \text{for all } a,$$

which implies $h_1 = h_2$. This proves the uniqueness of h.

Using the above, we make the following definition:

Definition 4.2. 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). Indeed, as $\mathbf{Oa} = \mathbf{Ob} + \mathbf{ba}$ and $b \in H \cap S^{n-1}$ i.e., $\mathbf{Oa} \cdot \mathbf{Ob} = 1$ and $\|\mathbf{Ob}\|^2 = 1$, we get

$$1 = \mathbf{Oa} \cdot \mathbf{Ob} = (\mathbf{Ob} + \mathbf{ba}) \cdot \mathbf{Ob} = \|\mathbf{Ob}\|^2 + \mathbf{ba} \cdot \mathbf{Ob} = 1 + \mathbf{ba} \cdot \mathbf{Ob},$$

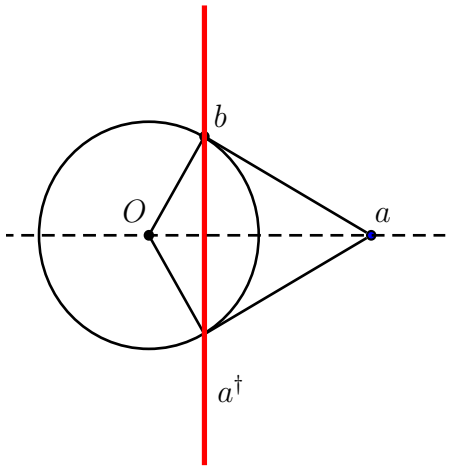


Figure 4.10: The polar, $a^\dagger$, of a point, a, outside the sphere S^{n-1}

which implies $\mathbf{ba} \cdot \mathbf{Ob} = 0$. When $a \in S^{n-1}$, the hyperplane $a^\dagger$ is tangent to S^{n-1} at a.

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. Indeed, $H^\dagger$ is the unique point so that

$$H = \{b \in \mathbb{E}^n \mid \mathbf{OH}^\dagger \cdot \mathbf{Ob} = 1\}$$

and

$$a^\dagger = \{b \in \mathbb{E}^n \mid \mathbf{Oa} \cdot \mathbf{Ob} = 1\};$$

since $a \in H$, we have $\mathbf{OH}^\dagger \cdot \mathbf{Oa} = 1$, which shows that $H^\dagger \in a^\dagger$.

If $a = (a_1, \ldots, a_n)$, the equation of the polar hyperplane $a^\dagger$ is

$$a_1 X_1 + \cdots + a_n X_n = 1.$$

Remark: As we noted, polarity in a Euclidean space suffers from the minor defect that the polar of the origin is undefined and, similarly, the pole of a hyperplane through the origin does not make sense. If we embed $\mathbb{E}^n$ into the projective space, $\mathbb{P}^n$, by adding a "hyperplane at infinity" (a copy of $\mathbb{P}^{n-1}$), thereby viewing $\mathbb{P}^n$ as the disjoint union $\mathbb{P}^n = \mathbb{E}^n \cup \mathbb{P}^{n-1}$, then the polarity correspondence can be defined everywhere. Indeed, the polar of the origin is the hyperplane at infinity ($\mathbb{P}^{n-1}$) and since $\mathbb{P}^{n-1}$ can be viewed as the set of hyperplanes through the origin in $\mathbb{E}^n$, the pole of a hyperplane through the origin is the corresponding "point at infinity" in $\mathbb{P}^{n-1}$.

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 4.3. 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.

For simplicity of notation, we write $a^\dagger_-$ for $(a^\dagger)_-$. Observe that $\{O\}^* = \mathbb{E}^n$, so it is convenient to set $O^\dagger_- = \mathbb{E}^n$, even though $O^\dagger$ is undefined. 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 $B^n(r)^* = B^n(1/r)$.

In Figure 4.11, 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. These polar lines are all easy to determine by drawing for each vertex, v_i, the tangent lines to the circle and joining the contact points. The construction of the polar of v_3 is shown in detail.

Remark: We chose a different notation for polar hyperplanes and polars ($a^\dagger$ and $H^\dagger$) and polar duals (A^*), to avoid the potential confusion between $H^\dagger$ and H^*, where H is a hyperplane (or $a^\dagger$ and $\{a\}^*$, where a is a point). Indeed, they are completely different! For example, the polar dual of a hyperplane is either a line orthogonal to H through O, if $O \in H$, or a semi-infinite line through O and orthogonal to H whose endpoint is the pole, $H^\dagger$, of H, whereas, $H^\dagger$ is a single point! Ziegler ([69], Chapter 2) use the notation $A^\triangle$ instead of A^* for the polar dual of A.

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 4.22. *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 $A^{**} = A$.

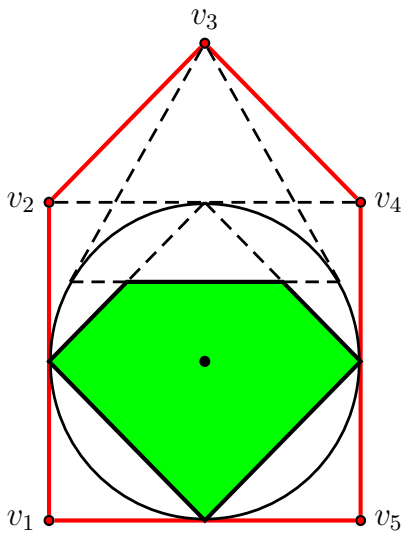


Figure 4.11: The polar dual of a polygon

Proof. (i) If A is bounded, then $A \subseteq B^n(r)$ for some $r > 0$ large enough. Then, $B^n(r)^* = B^n(1/r) \subseteq A^*$, so that $O \in \overset{\circ}{A}{}^*$. If $O \in \overset{\circ}{A}$, then $B^n(r) \subseteq A$ for some r small enough, so $A^* \subseteq B^n(r)^* = B^r(1/r)$ and A^* is bounded.

(ii) We always have $A \subseteq A^{**}$. We prove that if $b \notin A$, then $b \notin A^{**}$; this shows that $A^{**} \subseteq A$ and thus, $A = A^{**}$. Since A is closed and convex and $\{b\}$ is compact (and convex!), by Corollary 4.10, there is a hyperplane, H, strictly separating A and b and, in particular, $O \notin H$, as $O \in A$. If $h = H^\dagger$ is the pole of H, we have

$$\mathbf{Oh} \cdot \mathbf{Ob} > 1 \quad \text{and} \quad \mathbf{Oh} \cdot \mathbf{Oa} < 1, \quad \text{for all } a \in A$$

since $H_- = \{a \in \mathbb{E}^n \mid \mathbf{Oh} \cdot \mathbf{Oa} \leq 1\}$. This shows that $b \notin A^{**}$, since

$$\begin{aligned} A^{**} &= \{c \in \mathbb{E}^n \mid \mathbf{Od} \cdot \mathbf{Oc} \leq 1 \quad \text{for all } d \in A^*\} \\ &= \{c \in \mathbb{E}^n \mid (\forall d \in \mathbb{E}^n)(\text{if} \quad \mathbf{Od} \cdot \mathbf{Oa} \leq 1 \quad \text{for all } a \in A, \quad \text{then} \quad \mathbf{Od} \cdot \mathbf{Oc} \leq 1)\}, \end{aligned}$$

just let $c = b$ and $d = h$. □

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 4.22 will play a key role in studying polytopes, but before doing this, we need one more proposition.

Proposition 4.23. *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.*

Proof. Since $O \in \overset{\circ}{A}$, we have $O \notin \partial A$, and so, for every $a \in \partial A$, the polar hyperplane $a^\dagger$ is well-defined. Pick any $a \in \partial A$ and let $H = a^\dagger$ be its polar hyperplane. By definition, $A^* \subseteq H_-$, the closed half-space determined by H and containing O. If T is any supporting hyperplane to A at a, as $a \in T$, we have $t = T^\dagger \in a^\dagger = H$. Furthermore, it is a simple exercise to prove that $t \in (T_-)^*$ (in fact, $(T_-)^*$ is the interval with endpoints O and t). Since $A \subseteq T_-$ (because T is a supporting hyperplane to A at a), we deduce that $t \in A^*$, and thus, H is a supporting hyperplane to A^* at t. By Proposition 4.22, as A is closed and convex, $A^{**} = A$; it follows that all supporting hyperplanes to A^* are indeed obtained this way. $\square$

Chapter 5

Polyhedra and Polytopes

5.1 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 [69]).

Definition 5.1. 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, $p \geq 1$, 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, $P = \mathrm{conv}(S)$, of a finite set of points, $S \subseteq \mathcal{E}$.

[1] This means that the vector space, $\overrightarrow{\mathcal{E}}$, associated with $\mathcal{E}$ is a Euclidean space.

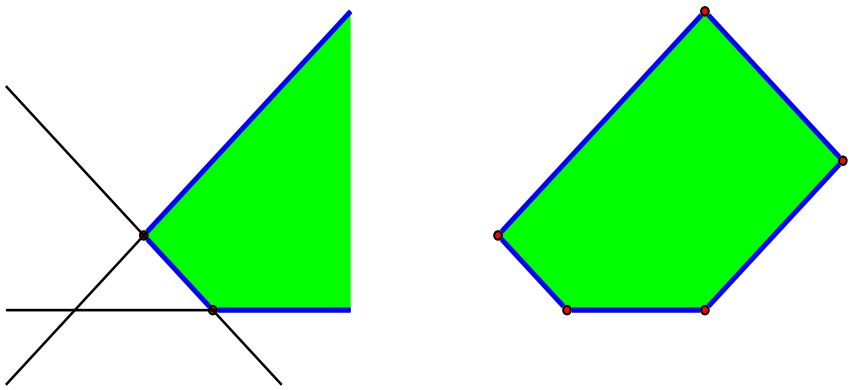


Figure 5.1: (a) An $\mathcal{H}$-polyhedron. (b) A $\mathcal{V}$-polytope

Obviously, $\mathcal{H}$-polyhedra are convex and closed as intersections of convex and closed half-spaces (in $\mathcal{E}$). By Proposition 3.3, $\mathcal{V}$-polytopes are also convex and closed (in $\mathcal{E}$). Since the notions of $\mathcal{H}$-polytope and $\mathcal{V}$-polytope are equivalent (see Theorem 5.7), we often use the simpler locution polytope. Examples of an $\mathcal{H}$-polyhedron and of a $\mathcal{V}$-polytope are shown in Figure 5.1.

Note that Definition 5.1 allows $\mathcal{H}$-polytopes and $\mathcal{V}$-polytopes to have an empty interior, which is somewhat of an inconvenience. This is not a problem, since we may always restrict ourselves to the affine hull of P (some affine space, E, of dimension $d \leq n$, where $d = \dim(P)$, as in Definition 3.2) as we now show.

Proposition 5.1. *Let $A \subseteq \mathcal{E}$ be a $\mathcal{V}$-polytope or an $\mathcal{H}$-polyhedron, let $E = \mathrm{aff}(A)$ be the affine hull of A in $\mathcal{E}$ (with the Euclidean structure on E induced by the Euclidean structure on $\mathcal{E}$) and write $d = \dim(E)$. Then, the following assertions hold:*

(1) The set, A, is a $\mathcal{V}$-polytope in E (i.e., viewed as a subset of E) iff A is a $\mathcal{V}$-polytope in $\mathcal{E}$.

(2) The set, A, is an $\mathcal{H}$-polyhedron in E (i.e., viewed as a subset of E) iff A is an $\mathcal{H}$-polyhedron in $\mathcal{E}$.

Proof. (1) This follows immediately because E is an affine subspace of $\mathcal{E}$ and every affine subspace of $\mathcal{E}$ is closed under affine combinations and so, *a fortiori*, under convex combinations. We leave the details as an easy exercise.

(2) Assume A is an $\mathcal{H}$-polyhedron in $\mathcal{E}$ and that $d < n$. By definition, $A = \bigcap_{i=1}^{p} C_i$, where the C_i are closed half-spaces determined by some hyperplanes, $H_1, \ldots, H_p$, in $\mathcal{E}$. (Observe that the hyperplanes, H_i's, associated with the closed half-spaces, C_i, may not be distinct.

For example, we may have $C_i = (H_i)_+$ and $C_j = (H_i)_-$, for the two closed half-spaces determined by H_i.) As $A \subseteq E$, we have

$$A = A \cap E = \bigcap_{i=1}^{p} (C_i \cap E),$$

where $C_i \cap E$ is one of the closed half-spaces determined by the hyperplane, $H_i' = H_i \cap E$, in E. Thus, A is also an $\mathcal{H}$-polyhedron in E.

Conversely, assume that A is an $\mathcal{H}$-polyhedron in E and that $d < n$. As any hyperplane, H, in $\mathcal{E}$ can be written as the intersection, $H = H_- \cap H_+$, of the two closed half-spaces that it bounds, E itself can be written as the intersection,

$$E = \bigcap_{i=1}^{p} E_i = \bigcap_{i=1}^{p} (E_i)_+ \cap (E_i)_-,$$

of finitely many half-spaces in $\mathcal{E}$. Now, as A is an $\mathcal{H}$-polyhedron in E, we have

$$A = \bigcap_{j=1}^{q} C_j,$$

where the C_j are closed half-spaces in E determined by some hyperplanes, H_j, in E. However, each H_j can be extended to a hyperplane, H_j', in $\mathcal{E}$, and so, each C_j can be extended to a closed half-space, C_j', in $\mathcal{E}$ and we still have

$$A = \bigcap_{j=1}^{q} C_j'.$$

Consequently, we get

$$A = A \cap E = \bigcap_{i=1}^{p} ((E_i)_+ \cap (E_i)_-) \cap \bigcap_{j=1}^{q} C_j',$$

which proves that A is also an $\mathcal{H}$-polyhedron in $\mathcal{E}$. □

The following simple proposition shows that we may assume that $\mathcal{E} = \mathbb{E}^n$:

Proposition 5.2. *Given any two affine Euclidean spaces, E and F, if $h\colon E \to F$ is any affine map then:*

(1) If A is any $\mathcal{V}$-polytope in E, then $h(E)$ is a $\mathcal{V}$-polytope in F.

(2) If h is bijective and A is any $\mathcal{H}$-polyhedron in E, then $h(E)$ is an $\mathcal{H}$-polyhedron in F.

Proof. (1) As any affine map preserves affine combinations it also preserves convex combination. Thus, $h(\mathrm{conv}(S)) = \mathrm{conv}(h(S))$, for any $S \subseteq E$.

(2) Say $A = \bigcap_{i=1}^{p} C_i$ in E. Consider any half-space, C, in E and assume that

$$C = \{x \in E \mid \varphi(x) \leq 0\},$$

for some affine form, φ, defining the hyperplane, $H = \{x \in E \mid \varphi(x) = 0\}$. Then, as h is bijective, we get

$$\begin{aligned} h(C) &= \{h(x) \in F \mid \varphi(x) \leq 0\} \\ &= \{y \in F \mid \varphi(h^{-1}(y)) \leq 0\} \\ &= \{y \in F \mid (\varphi \circ h^{-1})(y) \leq 0\}. \end{aligned}$$

This shows that $h(C)$ is one of the closed half-spaces in F determined by the hyperplane, $H' = \{y \in F \mid (\varphi \circ h^{-1})(y) = 0\}$. Furthermore, as h is bijective, it preserves intersections so

$$h(A) = h\left(\bigcap_{i=1}^{p} C_i\right) = \bigcap_{i=1}^{p} h(C_i),$$

a finite intersection of closed half-spaces. Therefore, $h(A)$ is an $\mathcal{H}$-polyhedron in F. □

By Proposition 5.2 we may assume that $\mathcal{E} = \mathbb{E}^d$ and by Proposition 5.1 we may assume that $\dim(A) = d$. These propositions justify the type of argument beginning with: "We may assume that $A \subseteq \mathbb{E}^d$ has dimension d, that is, that A has nonempty interior." This kind of reasonning will occur many times.

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 $p \times n$ 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 $b_i = 0$. This observation will be useful to show that the polar dual of an $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron.

Remark: Some authors call "convex" polyhedra and "convex" polytopes what we have simply called polyhedra and polytopes. Since Definition 5.1 implies that these objects are

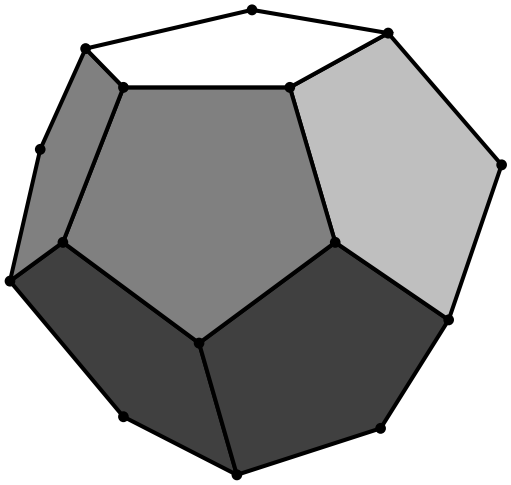


Figure 5.2: Example of a polytope (a dodecahedron)

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 [69], Berger [8], Grunbaum [36] and especially Cromwell [22], for pictures of polyhedra and polytopes. Figure 5.2 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. Figure 5.3 shows a polytope called an *icosahedron* whose faces are triangles. The icosahedron has 20 faces, 30 edges and 12 vertices.

Figure 5.3: Another Example of a polytope (an icosahedron)

Even better and a lot more entertaining, take a look at the spectacular web sites of George Hart,

Virtual Polyedra: http://www.georgehart.com/virtual-polyhedra/vp.html,

George Hart's web site: http://www.georgehart.com/

and also

Zvi Har'El's web site: http://www.math.technion.ac.il/ rl/

The *Uniform Polyhedra* web site: http://www.mathconsult.ch/showroom/unipoly/

Paper Models of Polyhedra: http://www.korthalsaltes.com/

Bulatov's *Polyhedra Collection*: http://www.physics.orst.edu/ bulatov/polyhedra/

Paul Getty's *Polyhedral Solids*: http://home.teleport.com/ tpgettys/poly.shtml

Jill Britton's *Polyhedra Pastimes*: http://ccins.camosun.bc.ca/ jbritton/jbpolyhedra.htm

and many other web sites dealing with polyhedra in one way or another by searching for "polyhedra" on *Google*!

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.

In order to prove that an $\mathcal{H}$-polytope is a $\mathcal{V}$-polytope we need to take a closer look at polyhedra. Basically, we need to make precise the notion of vertex, edge, and face, which are intuitively clear in dimension three.

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). The following proposition shows that the C_i in an irredundant decomposition of A are uniquely determined by A.

Proposition 5.3. *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 $i \neq j$.*

Proof. (ii) Fix any i and consider $A_i = \bigcap_{j \neq i} C_j$. As $A = \bigcap_{i=1}^{t} C_i$ is an irredundant decomposition, there is some $x \in A_i - C_i$. Pick any $a \in \overset{\circ}{A}$. By Lemma 4.1, we get $b = [a, x] \cap H_i \in \overset{\circ}{A}_i$, so b belongs to the interior of $H_i \cap A_i$ in H_i.

(iii) As $\partial A = A - \overset{\circ}{A} = A \cap (\overset{\circ}{A})^c$ (where B^c denotes the complement of a subset B of $\mathbb{E}^n$) and $\partial C_i = H_i$, we get

$$\begin{aligned}
\partial A &= \left(\bigcap_{i=1}^{t} C_i\right) - \left(\overset{\circ}{\bigcap_{j=1}^{t}} C_j\right) \\
&= \left(\bigcap_{i=1}^{t} C_i\right) - \left(\bigcap_{j=1}^{t} \overset{\circ}{C}_j\right) \\
&= \left(\bigcap_{i=1}^{t} C_i\right) \cap \left(\bigcap_{j=1}^{t} \overset{\circ}{C}_j\right)^c \\
&= \left(\bigcap_{i=1}^{t} C_i\right) \cap \left(\bigcup_{j=1}^{t} (\overset{\circ}{C}_j)^c\right) \\
&= \bigcup_{j=1}^{t} \left(\left(\bigcap_{i=1}^{t} C_i\right) \cap (\overset{\circ}{C}_j)^c \right) \\
&= \bigcup_{j=1}^{t} \left(\partial C_j \cap \left(\bigcap_{i \neq j} C_i\right) \right) \\
&= \bigcup_{j=1}^{t} (H_j \cap A) = \bigcup_{j=1}^{t} \mathrm{Facet}_j\, A.
\end{aligned}$$

If we had $\mathrm{Facet}_i\, A \subseteq \mathrm{Facet}_j\, A$, for some $i \neq j$, then, by (ii), as the affine hull of $\mathrm{Facet}_i\, A$ is H_i and the affine hull of $\mathrm{Facet}_j\, A$ is H_j, we would have $H_i \subseteq H_j$, a contradiction.

(i) As the decomposition is irredundant, the H_i are pairwise distinct. Also, by (ii), each facet, $\mathrm{Facet}_i\, A$, has dimension $d-1$ (where $d = \dim A$). Then, in (iii), we can show that the decomposition of ∂A as a union of polytopes of dimension $d-1$ whose pairwise nonempty intersections have dimension at most $d-2$ (since they are contained in pairwise distinct hyperplanes) is unique up to permutation. Indeed, assume that

$$\partial A = F_1 \cup \cdots \cup F_m = G_1 \cup \cdots \cup G_n,$$

where the F_i's and G'_j are polyhedra of dimension $d-1$ and each of the unions is irredundant. Then, we claim that for each F_i, there is some $G_{\varphi(i)}$ such that $F_i \subseteq G_{\varphi(i)}$. If not, F_i would be expressed as a union

$$F_i = (F_i \cap G_{i_1}) \cup \cdots \cup (F_i \cap G_{i_k})$$

where $\dim(F_i \cap G_{i_j}) \leq d-2$, since the hyperplanes containing F_i and the G_j's are pairwise distinct, which is absurd, since $\dim(F_i) = d-1$. By symmetry, for each G_j, there is some $F_{\psi(j)}$ such that $G_j \subseteq F_{\psi(j)}$. But then, $F_i \subseteq F_{\psi(\varphi(i))}$ for all i and $G_j \subseteq G_{\varphi(\psi(j))}$ for all j which implies $\psi(\varphi(i)) = i$ for all i and $\varphi(\psi(j)) = j$ for all j since the unions are irredundant. Thus, φ and ψ are mutual inverses and the B_j's are just a permutation of the A_i's, as claimed. Therefore, the facets, $\mathrm{Facet}_i A$, are uniquely determined by A and so are the hyperplanes, $H_i = \mathrm{aff}(\mathrm{Facet}_i A)$, and the half-spaces, C_i, that they determine. □

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 5.2. 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 3.6. This is true and, in fact, the vertices of a polyhedron coincide with its extreme points (see Definition 3.7).

Proposition 5.4. *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.

Proof. (1) If H is a supporting hyperplane to A at a, then, one of the half-spaces, C, determined by H, satisfies $A = A \cap C$. It follows from Proposition 5.3 that if $H \neq H_i$ (where

[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).

the hyperplanes H_i are the supporting hyperplanes of the facets of A), then C is redundant, from which (1) follows.

(2) If $a \in \partial A$ is not extreme, then $a \in [y, z]$, where $y, z \in \partial A$. However, this implies that a has order $k \geq 1$, i.e, a is not a vertex. □

The proof that every $\mathcal{H}$-polytope A is a $\mathcal{V}$-polytope relies on the fact that the extreme points of an $\mathcal{H}$-polytope coincide with its vertices, which form a finite nonempty set, and by Krein and Millman's Theorem (Theorem 3.8), A is the convex hull of its vertices.

The proof that every $\mathcal{V}$-polytope A is an $\mathcal{H}$-polytope relies on the crucial fact that the polar dual A^* of a $\mathcal{V}$-polytope A is an $\mathcal{H}$-polyhedron, and that the equations of the hyperplanes cutting out A^* are obtained in a very simple manner from the points a_i specifying A as $A = \text{conv}(a_1, \ldots, a_p)$; see Proposition 5.5.

The proof that every $\mathcal{V}$-polytope A is an $\mathcal{H}$-polytope consists of the following steps:

(1) Construct the polar dual A^* of A. Then we know that A^* is is an $\mathcal{H}$-polyhedron. Furthermore, if the center O of the polar duality belongs to the interior of A, then A^* is bounded, and so it is an $\mathcal{H}$-polytope.

(2) Since A^* is $\mathcal{H}$-polytope, A^* is also a $\mathcal{V}$-polytope, as we claimed earlier.

(3) Since A^* is a $\mathcal{V}$-polytope, its polar dual A^{**} is an $\mathcal{H}$-polyhedron.

(4) A $\mathcal{V}$-polytope is closed and convex, and since O belong to A (in fact, to the interior of A), by Proposition 4.22, we have $A = A^{**}$, so A is indeed an $\mathcal{H}$-polyhedron; in fact, A an $\mathcal{H}$-polytope since it is bounded.

5.2 Polar Duals of $\mathcal{V}$-Polytopes and $\mathcal{H}$-Polyhedra of the Form $P(A, \mathbf{1})$

The following proposition gives a simple description of the polar dual A^* of a $\mathcal{V}$-polyhedron $A = \text{conv}(v_1, \ldots, v_p)$ in terms of the hyperplanes $a_i^\dagger$. It is a key ingredient in the proof of the equivalence of $\mathcal{V}$-polytopes and $\mathcal{H}$-polytopes because it implies that if the center of polar duality is chosen in the interior of a $\mathcal{V}$-polytope, then its poar dual is an $\mathcal{H}$-polytope.

Proposition 5.5. *Let $S = \{a_i\}_{i=1}^p$ be a finite set of points in $\mathbb{E}^n$ and let $A = \text{conv}(S)$ be its convex hull. If $S \neq \{O\}$, then, the dual A^* of A w.r.t. the center O is the $\mathcal{H}$-polyhedron given by*

$$A^* = \bigcap_{i=1}^{p} (a_i^\dagger)_-.$$

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 $A^* = \mathbb{E}^d$.

Proof. By definition, we have

$$A^* = \{b \in \mathbb{E}^n \mid \mathbf{Ob} \cdot (\sum_{j=1}^{p} \lambda_j \mathbf{Oa_j}) \leq 1, \quad \lambda_j \geq 0, \sum_{j=1}^{p} \lambda_j = 1\},$$

and the right hand side is clearly equal to $\bigcap_{i=1}^{p} \{b \in \mathbb{E}^n \mid \mathbf{Ob} \cdot \mathbf{Oa_i} \leq 1\} = \bigcap_{i=1}^{p} (a_i^\dagger)_-$, which is a polyhedron. (Recall that $(a_i^\dagger)_- = \mathbb{E}^n$ if $a_i = O$.) If $O \in \overset{\circ}{A}$, then A^* is bounded (by Proposition 4.22) and so, A^* is an $\mathcal{H}$-polytope. □

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.

It is convenient to restate Proposition 5.5 using matrices. First, observe that the proof of Proposition 5.5 shows that

$$\text{conv}(\{a_1, \ldots, a_p\})^* = \text{conv}(\{a_1, \ldots, a_p\} \cup \{O\})^*.$$

Therefore, we may assume that not all $a_i = O$ $(1 \leq i \leq p)$. 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 $\text{conv}(\{a_1, \ldots, a_p\})^*$ can be written in matrix form as

$$\text{conv}(\{a_1, \ldots, a_p\})^* = \{x \in \mathbb{R}^n \mid A^\top x \leq \mathbf{1}\},$$

where $\mathbf{1}$ denotes the vector of $\mathbb{R}^p$ with all coordinates equal to 1. We write $P(A^\top, \mathbf{1}) = \{x \in \mathbb{R}^n \mid A^\top x \leq \mathbf{1}\}$. There is a useful converse of this property as proved in the next proposition.

Proposition 5.6. *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*

$$\text{conv}(\{a_1, \ldots, a_p\})^* = P(A^\top, \mathbf{1}),$$

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

$$P(A, \mathbf{1})^* = \text{conv}(\{a_1, \ldots, a_p\} \cup \{0\}),$$

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.

Proof. Only the second part needs a proof. Let $B = \text{conv}(\{a_1, \ldots, a_p\} \cup \{0\})$, where $a_i \in \mathbb{R}^n$ is the i^{th} *row* of A. Then, by the first part,

$$B^* = P(A, \mathbf{1}).$$

As $0 \in B$ and B is convex and closed (by Proposition 3.3), by Proposition 4.22, $B = B^{**} = P(A, \mathbf{1})^*$, as claimed. □

Remark: Proposition 5.6 still holds if A is the zero matrix because then, the inequalities $A^\top x \leq \mathbf{1}$ (or $Ax \leq \mathbf{1}$) are trivially satisfied. In the first case, $P(A^\top, \mathbf{1}) = \mathbb{E}^d$, and in the second case, $P(A, \mathbf{1}) = \mathbb{E}^d$.

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.

It is not clear that every $\mathcal{H}$-polyhedron is of the form $P(A, \mathbf{1})$. This is indeed the case if we pick O in the interior of A, but this is nontrivial to prove. What we will need is to find the corresponding "$\mathcal{V}$-definition" of an $\mathcal{H}$-polyhedron. For this we will need to add positive combinations of vectors to convex combinations of points. Intuitively, these vectors correspond to "points at infinity."

5.3 The Equivalence of $\mathcal{H}$-Polytopes and $\mathcal{V}$-Polytopes

We are now ready for the theorem showing the equivalence of $\mathcal{V}$-polytopes and $\mathcal{H}$-polytopes. This is a nontrivial theorem usually attributed to Weyl and Minkowski (for example, see Barvinok [4]).

Theorem 5.7. *(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.*

Proof. By restricting ourselves to the affine hull of A (some $\mathbb{E}^d$, with $d \leq n$) we may assume that A has nonempty interior. Since an $\mathcal{H}$-polytope has finitely many facets, we deduce by induction that an $\mathcal{H}$-polytope has a finite number of k-faces, for $k = 0, \ldots, d-2$. In particular, an $\mathcal{H}$-polytope has finitely many vertices. By proposition 5.4, these vertices are the extreme points of A and since an $\mathcal{H}$-polytope is compact and convex, by the theorem of Krein and Milman (Theorem 3.8), A is the convex hull of its set of vertices.

Conversely, again, we may assume that A has nonempty interior by restricting ourselves to the affine hull of A. Then, pick an origin O in the interior of A and consider the dual A^* of A. By Proposition 5.5, the convex set A^* is an $\mathcal{H}$-polytope. By the first part of the proof of Theorem 5.7, the $\mathcal{H}$-polytope A^* is the convex hull of its vertices. Finally, since A is convex and closed (by Proposition 3.3), and since O is in the interior of A, Proposition 4.22 and Proposition 5.5 hold, and we deduce that $A = A^{**}$ is an $\mathcal{H}$-polytope. □

In view of Theorem 5.7, we are justified in dropping the $\mathcal{V}$ or $\mathcal{H}$ in front of polytope, and will do so from now on. Theorem 5.7 has some interesting corollaries regarding the dual of a polytope.

Corollary 5.8. *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 5.9. *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.*

Proof. Immediate from Proposition 5.4 and Proposition 4.23. □

We also have the following proposition whose proof would not be that simple if we only had the notion of an $\mathcal{H}$-polytope (as a matter of fact, there is a way of proving Theorem 5.7 using Proposition 5.10).

Proposition 5.10. *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 $\mathbb{E}^m$.*

Proof. Immediate, since an $\mathcal{H}$-polytope is a $\mathcal{V}$-polytope and since affine maps send convex sets to convex sets. □

The reader should check that the Minkowski sum of polytopes is a polytope.

We were able to give a short proof of Theorem 5.7 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 Proposition 5.21, and also Ziegler [69] (Chapter 1) and Section 5.4. 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 and to generalize the equivalence theorem to $\mathcal{H}$-polyhedra and $\mathcal{V}$-polyhedra.

5.4 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 closed half-spaces that are not necessarily bounded. This equivalence was first proved by Motzkin in the early 1930's. It can be proved in several ways, some involving cones.

Definition 5.3. Let $\mathcal{E}$ be any affine Euclidean space of finite dimension n (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 nonnegative scalars called *positive combinations*. Given a subset, $V \subseteq \overrightarrow{\mathcal{E}}$, the *conical hull* or *positive hull* of V is the set

$$\mathrm{cone}(V) = \Big\{ \sum_I \lambda_i v_i \mid \{v_i\}_{i \in I} \subseteq V,\ \lambda_i \geq 0 \quad \text{for all } i \in I \Big\}.$$

A $\mathcal{V}$*-polyhedron* or *polyhedral set* is a subset $A \subseteq \mathcal{E}$ such that

$$A = \mathrm{conv}(Y) + \mathrm{cone}(V) = \{a + v \mid a \in \mathrm{conv}(Y),\ v \in \mathrm{cone}(V)\},$$

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,

$$C = \mathrm{cone}(\{u_1, \ldots, u_p\}),$$

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 $\mathrm{cone}(V)$ of V is also denoted $\mathrm{pos}(V)$. 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, this should not cause any confusion. In this section, we favor the view that $\mathcal{V}$-cones are special kinds of $\mathcal{V}$-polyhedra. As a consequence, a ($\mathcal{V}$ or $\mathcal{H}$)-cone always contains 0, sometimes called an *apex* of the cone.

A set of the form $\{a + tu \mid t \geq 0\}$, where $a \in \mathcal{E}$ is a point and $u \in \overrightarrow{\mathcal{E}}$ is a nonzero vector, is called a *half-line* or *ray*. Then, we see that a $\mathcal{V}$-polyhedron, $A = \mathrm{conv}(Y) + \mathrm{cone}(V)$, is the convex hull of the union of a finite set of points with a finite set of rays. In the case of a $\mathcal{V}$-cone, all these rays meet in a common point, an apex of the cone.

Since an $\mathcal{H}$-polyhedron is an intersection of half-spaces determined by hyperplanes, and since half-spaces are closed, an $\mathcal{H}$-polyhedron is closed. We know from Proposition 3.3 that a $\mathcal{V}$-polytope is closed and by Proposition 4.13 that a $\mathcal{V}$-cone is closed. To apply Proposition 4.22 to an arbitrary $\mathcal{V}$-polyhedron we need to know that a $\mathcal{V}$-polyhedron is closed.

Given a $\mathcal{V}$-polyhedron $P = \mathrm{conv}(Y) + \mathrm{cone}(V)$ of dimension d, an easy way to prove that P is closed is to "lift" P to the hyperplane H_{d+1} of equation $x_{d+1} = 1$ in $\mathbb{A}^{d+1}$, obtaining a polyhedron $\widehat{P}$ contained in H_{d+1} homeomorphic to P, and to consider a polyhedral cone (a $\mathcal{V}$-cone) $C(P)$ associated with P which has the property that

$$\widehat{P} = C(P) \cap H_{d+1}.$$

The details of this construction are given in Section 5.5; see Proposition 5.20(2). Since by Proposition 4.13 a $\mathcal{V}$-cone is closed and since a hyperplane is closed, $\widehat{P} = C(P) \cap H_{d+1}$ is closed, and thus P is closed. As a summary, the following propostion holds.

Proposition 5.11. *Every $\mathcal{V}$-polyhedron $P = \mathrm{conv}(Y) + \mathrm{cone}(V)$ is closed.*

Propositions 5.1 and 5.2 generalize easily to $\mathcal{V}$-polyhedra and cones.

Proposition 5.12. *Let $A \subseteq \mathcal{E}$ be a $\mathcal{V}$-polyhedron or an $\mathcal{H}$-polyhedron, let $E = \mathrm{aff}(A)$ be the affine hull of A in $\mathcal{E}$ (with the Euclidean structure on E induced by the Euclidean structure on $\mathcal{E}$) and write $d = \dim(E)$. Then, the following assertions hold:*

(1) The set A is a $\mathcal{V}$-polyhedron in E (i.e., viewed as a subset of E) iff A is a $\mathcal{V}$-polyhedron in $\mathcal{E}$.

(2) The set A is an $\mathcal{H}$-polyhedron in E (i.e., viewed as a subset of E) iff A is an $\mathcal{H}$-polyhedron in $\mathcal{E}$.

Proof. We already proved (2) in Proposition 5.1. For (1), observe that the direction $\overrightarrow{E}$ of E is a linear subspace of $\overrightarrow{\mathcal{E}}$. Consequently, E is closed under affine combinations and $\overrightarrow{E}$ is closed under linear combinations and the result follows immediately. □

Proposition 5.13. *Given any two affine Euclidean spaces E and F, if $h\colon E \to F$ is any affine map then:*

(1) If A is any $\mathcal{V}$-polyhedron in E, then $h(E)$ is a $\mathcal{V}$-polyhedron in F.

(2) If $g\colon \overrightarrow{E} \to \overrightarrow{F}$ is any linear map and if C is any $\mathcal{V}$-cone in $\overrightarrow{E}$, then $g(C)$ is a $\mathcal{V}$-cone in $\overrightarrow{F}$.

(3) If h is bijective and A is any $\mathcal{H}$-polyhedron in E, then $h(E)$ is an $\mathcal{H}$-polyhedron in F.

Proof. We already proved (3) in Proposition 5.2. For (1), using the fact that $h(a + u) = h(a) + \overrightarrow{h}(u)$ for any affine map, h, where $\overrightarrow{h}$ is the linear map associated with h, we get

$$h(\mathrm{conv}(Y) + \mathrm{cone}(V)) = \mathrm{conv}(h(Y)) + \mathrm{cone}(\overrightarrow{h}(V)).$$

For (2), as g is linear, we get

$$g(\mathrm{cone}(V)) = \mathrm{cone}(g(V)),$$

establishing the proposition. □

Propositions 5.12 and 5.13 allow us to assume that $\mathcal{E} = \mathbb{E}^d$ and that our ($\mathcal{V}$ or $\mathcal{H}$)-polyhedra, $A \subseteq \mathbb{E}^d$, have nonempty interior (*i.e.* $\dim(A) = d$).

The generalization of Theorem 5.7 is that every $\mathcal{V}$-polyhedron A is an $\mathcal{H}$-polyhedron and conversely.

At first glance, it may seem that there is a small problem when $A = \mathbb{E}^d$. Indeed, Definition 5.3 allows the possibility that $\text{cone}(V) = \mathbb{E}^d$ for some finite subset, $V \subseteq \mathbb{R}^d$. This is because it is possible to generate a basis of $\mathbb{R}^d$ using finitely many positive combinations. On the other hand the definition of an $\mathcal{H}$-polyhedron, A, (Definition 5.1) assumes that $A \subseteq \mathbb{E}^n$ is cut out by *at least one* hyperplane. So, A is always contained in some half-space of $\mathbb{E}^n$ and strictly speaking, $\mathbb{E}^n$ is not an $\mathcal{H}$-polyhedron! The simplest way to circumvent this difficulty is to decree that $\mathbb{E}^n$ itself is a polyhedron, which we do.

Yet another solution is to assume that, unless stated otherwise, every finite set of vectors V that we consider when defining a polyhedron has the property that there is some hyperplane H through the origin so that all the vectors in V lie in only one of the two closed half-spaces determined by H. But then, the polar dual of a polyhedron can't be a single point! Therefore, we stick to our decision that $\mathbb{E}^n$ itself is a polyhedron.

To prove the equivalence of $\mathcal{H}$-polyhedra and $\mathcal{V}$-polyhedra, 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" [69] (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 later.

However, it turns out that the double dualization technique used in the proof of Theorem 5.7 can be easily adapted to prove that every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron. This is because if O belongs to the interior of the $\mathcal{V}$-polyhedron A, then its polar dual A^* is an $\mathcal{H}$-polytope; see Proposition 5.14. Then, just as in the proof of Theorem 5.7, we can use the theorem of Krein and Millman to show that A^* is a $\mathcal{V}$-polytope. By taking the polar dual of A^*, we obtain the fact that $A^{**} = A$ is an $\mathcal{H}$-polyhedron.

Moreover, the dual of an $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron; see Proposition 5.15. This fact can be used to prove that every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron by using the fact already shown that every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron!

Consequently 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; see Proposition 5.21.

The generalization of Proposition 5.5 to polyhedral sets is shown below. As before, the center of our polar duality is denoted by O. It is taken as the origin of $\mathbb{E}^d$. The new ingredient is that because a $\mathcal{V}$-polyhderon is defined by points and vectors, its polar dual is still cut out by hyperplanes, but the hyperplanes corresponding to vectors pass through the origin. To show this we need to define the "polar hyperplane" $u^\dagger$ of a vector u.

Definition 5.4. 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 $u^\dagger$ through O normal to u and on the "opposite side" of u.

The following proposition generalizing Proposition 5.5 is inspired by Exercise 7 in Chapter 3 of Grunbaum [36].

Proposition 5.14. *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 O, if $A \neq \{O\}$, then the polar dual A^* of A w.r.t. O 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 O belongs to the interior of A, then A^* is bounded, that is, A^* is an $\mathcal{H}$-polytope. If $A = \{O\}$, then A^* is the special polyhedron $A^* = \mathbb{E}^d$.

Proof. By definition of A^* w.r.t. O, we have

$$\begin{aligned} A^* &= \left\{ x \in \mathbb{E}^d \;\middle|\; \mathbf{Ox} \cdot \mathbf{O}\left(\sum_{\mathbf{i=1}}^{\mathbf{p}} \lambda_\mathbf{i} \mathbf{y_i} + \sum_{\mathbf{j=1}}^{\mathbf{q}} \mu_\mathbf{j} \mathbf{v_j}\right) \leq 1,\ \lambda_i \geq 0,\ \sum_{i=1}^{p} \lambda_i = 1,\ \mu_j \geq 0 \right\} \\ &= \left\{ x \in \mathbb{E}^d \;\middle|\; \sum_{i=1}^{p} \lambda_i \mathbf{Ox} \cdot \mathbf{Oy_i} + \sum_{j=1}^{q} \mu_j \mathbf{Ox} \cdot v_j \leq 1,\ \lambda_i \geq 0,\ \sum_{i=1}^{p} \lambda_i = 1,\ \mu_j \geq 0 \right\}. \end{aligned}$$

When $\mu_j = 0$ for $j = 1, \ldots, q$, we get

$$\sum_{i=1}^{p} \lambda_i \mathbf{Ox} \cdot \mathbf{Oy_i} \leq 1, \quad \lambda_i \geq 0,\ \sum_{i=1}^{p} \lambda_i = 1$$

and we check that

$$\begin{aligned} \left\{ x \in \mathbb{E}^d \;\middle|\; \sum_{i=1}^{p} \lambda_i \mathbf{Ox} \cdot \mathbf{Oy_i} \leq 1,\ \lambda_i \geq 0,\ \sum_{i=1}^{p} \lambda_i = 1 \right\} &= \bigcap_{i=1}^{p} \{x \in \mathbb{E}^d \mid \mathbf{Ox} \cdot \mathbf{Oy_i} \leq 1\} \\ &= \bigcap_{i=1}^{p} (y_i^\dagger)_-. \end{aligned}$$

The points in A^* must also satisfy the conditions

$$\sum_{j=1}^{q} \mu_j \mathbf{Ox} \cdot v_j \leq 1 - \alpha, \quad \mu_j \geq 0,\ \mu_j > 0 \text{ for some } j,\ 1 \leq j \leq q,$$

with $\alpha \leq 1$ (here $\alpha = \sum_{i=1}^{p} \lambda_i \mathbf{Ox} \cdot \mathbf{Oy_i}$). In particular, for every $j \in \{1, \ldots, q\}$, if we set $\mu_k = 0$ for $k \in \{1, \ldots, q\} - \{j\}$, we should have

$$\mu_j \mathbf{Ox} \cdot v_j \leq 1 - \alpha \quad \text{for all} \quad \mu_j > 0,$$

that is,

$$\mathbf{Ox} \cdot v_j \leq \frac{1-\alpha}{\mu_j} \quad \text{for all} \quad \mu_j > 0,$$

which is equivalent to

$$\mathbf{Ox} \cdot v_j \leq 0.$$

Consequently, if $x \in A^*$, we must also have

$$x \in \bigcap_{j=1}^{q} \{x \in \mathbb{E}^d \mid \mathbf{Ox} \cdot v_j \leq 0\} = \bigcap_{j=1}^{q} (v_j^\dagger)_-.$$

Therefore,

$$A^* \subseteq \bigcap_{i=1}^{p} (y_i^\dagger)_- \cap \bigcap_{j=1}^{q} (v_j^\dagger)_-.$$

However, the reverse inclusion is obvious and thus, we have equality. If O belongs to the interior of A, we know from Proposition 4.22 that A^* is bounded. Therefore, A^* is indeed an $\mathcal{H}$-polytope of the above form. □

It is fruitful to restate Proposition 5.14 in terms of matrices (as we did for Proposition 5.5). First, observe that

$$(\mathrm{conv}(Y) + \mathrm{cone}(V))^* = (\mathrm{conv}(Y \cup \{O\}) + \mathrm{cone}(V))^*.$$

If we pick O as an origin then we can represent the points in Y as vectors, and O is now denoted 0. The zero vector is denoted $\mathbf{0}$.

If $A = \mathrm{conv}(Y) + \mathrm{cone}(V) \neq \{0\}$, let Y be the $d \times p$ matrix whose i^{th} column is y_i and let V is the $d \times q$ matrix whose j^{th} column is v_j. Then Proposition 5.14 says that

$$(\mathrm{conv}(Y) + \mathrm{cone}(V))^* = \{x \in \mathbb{R}^d \mid Y^\top x \leq \mathbf{1},\ V^\top x \leq \mathbf{0}\}.$$

We write $P(Y^\top, \mathbf{1}; V^\top, \mathbf{0}) = \{x \in \mathbb{R}^d \mid Y^\top x \leq \mathbf{1},\ V^\top x \leq \mathbf{0}\}$.

If $A = \mathrm{conv}(Y) + \mathrm{cone}(V) = \{0\}$, then both Y and V must be zero matrices but then, the inequalities $Y^\top x \leq \mathbf{1}$ and $V^\top x \leq \mathbf{0}$ are trivially satisfied by all $x \in \mathbb{E}^d$, so even in this case we have

$$\mathbb{E}^d = (\mathrm{conv}(Y) + \mathrm{cone}(V))^* = P(Y^\top, \mathbf{1}; V^\top, \mathbf{0}).$$

The converse of Proposition 5.14 also holds as shown below.

Proposition 5.15. *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*

$$(\text{conv}(\{y_1, \ldots, y_p\}) + \text{cone}(\{v_1, \ldots, v_q\}))^* = P(Y^\top, \mathbf{1}; V^\top, \mathbf{0}),$$

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

$$P(Y, \mathbf{1}; V, \mathbf{0})^* = \text{conv}(\{y_1, \ldots, y_p\} \cup \{0\}) + \text{cone}(\{v_1, \ldots, v_q\}),$$

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,

$$P(Y, \mathbf{1}; V, \mathbf{0})^* = \{x \in \mathbb{R}^d \mid x = Y^\top u + V^\top t,\ u \in \mathbb{R}^p,\ t \in \mathbb{R}^q,\ u, t \geq 0,\ \mathbb{I}u = 1\},$$

where $\mathbb{I}$ is the row vector of length p whose coordinates are all equal to 1.

Proof. Only the second part needs a proof. Let

$$B = \text{conv}(\{y_1, \ldots, y_p\} \cup \{0\}) + \text{cone}(\{v_1, \ldots, v_q\}),$$

where $y_i \in \mathbb{R}^p$ is the i^{th} *row* of Y and $v_j \in \mathbb{R}^q$ is the j^{th} *row* of V. Then, by the first part,

$$B^* = P(Y, \mathbf{1}; V, \mathbf{0}).$$

As $0 \in B$ and B is closed (by Proposition 5.11) and convex, by Proposition 4.22, $B = B^{**} = P(Y, \mathbf{1}; V, \mathbf{0})^*$, as claimed. □

Proposition 5.15 has the following important corollary:

Proposition 5.16. *The following assertions hold:*

(1) The polar dual A^* of every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron.

(2) The polar dual A^* of every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron.

Proof. (1) We may assume that $0 \in A$, in which case A is of the form $A = P(Y, \mathbf{1}; V, \mathbf{0})$. By the second part of Proposition 5.15, A^* is a $\mathcal{V}$-polyhedron.

(2) This is the first part of Proposition 5.15. □

We can now use Proposition 5.14, Proposition 4.22, and Krein and Milman's Theorem to prove that every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron.

Proposition 5.17. *If $A \neq \mathbb{E}^d$ is a $\mathcal{V}$-polyhedron and if we choose the center of the polarity O in the interior $\overset{\circ}{A}$ of A, then A is of the form $A = P(Y, \mathbf{1})$. Therefore, every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron.*

Proof. Let $A \neq \mathbb{E}^d$ be a $\mathcal{V}$-polyhedron of dimension d. Thus $A \subseteq \mathbb{E}^d$ has nonempty interior so we can pick some point O in the interior of A. If $d = 0$, then $A = \{0\} = \mathbb{E}^0$ and we are done. Otherwise, by Proposition 5.14, the polar dual A^* of A w.r.t. O is an $\mathcal{H}$-polytope. Then, as in the proof of Theorem 5.7, using Krein and Milman's Theorem we deduce that A^* is a $\mathcal{V}$-polytope. Now, as O belongs to A and A is closed (by Proposition 5.11) and convex, by Proposition 4.22 (even if A is not bounded) we have $A = A^{**}$, and by Proposition 5.6 (first part), we conclude that $A = A^{**}$ is an $\mathcal{H}$-polyhedron 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 5.18. *Every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron.*

Proof. Let A be an $\mathcal{H}$-polyhedron of dimension d. The case where A is bounded is covered by Theorem 5.7 (Weyl-Minkowski), If A is unbounded, there are two cases to consider.

(1) The pohyhedron A is cut out by hyperplanes not containing 0, that is, $0 \in \overset{\circ}{A}$ and A is of the form $A = P(Y, \mathbf{1})$. In this case, by Proposition 5.6, A^* is a $\mathcal{V}$-polytope, but since A is unbounded 0 lies on the boundary of A^*. We can translate 0 using some translation Ω so that the new origin Ω is now in the interior of A^*, so by Theorem 5.7 (with respect to the origin Ω), A^* is an $\mathcal{H}$-polytope. We now translate Ω back to 0 using the translation $-\Omega$, but then 0 lies on the boundary of A^*, so some of the hyperplanes cutting out A^* contain 0, which implies that A^* is an $\mathcal{H}$-polyhedron of the form $A^* = P(Y, \mathbf{1}; V, \mathbf{0})$. By Proposition 5.15, we deduce that $A^{**} = A$ is a $\mathcal{V}$-polyhedron ($A = A^{**}$ because $0 \in A$ and A is closed and convex).

(2) The pohyhedron A is cut out by hyperplanes, some of which contain 0, which means A is of the form $A = P(Y, \mathbf{1}; V, \mathbf{0})$. By Proposition 5.15, the polar dual A^* of A is a $\mathcal{V}$-polyhedron. As in the previous case, 0 lies on the boundary of A^*. We translate 0 using some translation Ω so that the new origin Ω is now in the interior of A^*, and by Proposition 5.17, A^* is an $\mathcal{H}$-polyhedron. As in Case (1) we translate Ω back to 0 using the translation $-\Omega$, so A^* is of the form $A^* = P(Y, \mathbf{1}; V, \mathbf{0})$. By Proposition 5.15, we deduce that $A^{**} = A$ is a $\mathcal{V}$-polyhedron ($A = A^{**}$ because $0 \in A$ and A is closed and convex).

This concludes the proof that every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron. □

Putting together Propositions 5.17 and 5.18 we obtain our main theorem:

Theorem 5.19. *(Equivalence of $\mathcal{H}$-polyhedra and $\mathcal{V}$-polyhedra) Every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron and conversely.*

Both in Proposition 5.17 and in Proposition 5.18, the step that is not automatic is to find the vertices of an $\mathcal{H}$-polytope from the inequalities defining this $\mathcal{H}$-polytope. A method to do this algorithmically is Fourier–Motzkin elimination; see Proposition 5.21. This process

can be expansive, in the sense that the number of vertices can be exponential in the number of inequalities. For example, the standard d-cube is cut out by $2d$ hyperplanes, but is has 2^d vertices.

Here are some examples illustrating Proposition 5.17.

Example 5.1. Let A be the $\mathcal{V}$-polyhedron (a triangle) in $\mathbb{A}^2$ defined by set $Y = \{(-1, -1/2), (1, -1/2), (0, 1/2)\}$. By Proposition 5.14, the polar dual A^* is the $\mathcal{H}$-polytope, a triangle, cut out by the inequalities:

$$\begin{aligned} -x - (1/2)y &\leq 1 \\ x - (1/2)y &\leq 1 \\ (1/2)y &\leq 1. \end{aligned}$$

This is also the $\mathcal{V}$-polytope whose vertices are $(-2, 2), (2, 2), (0, -2)$. By Proposition 5.6, $A = A^{**}$ is $\mathcal{H}$-polyhedron cut out by the inequalities

$$\begin{aligned} -2x + 2y &\leq 1 \\ 2x + 2y &\leq 1 \\ -2y &\leq 1, \end{aligned}$$

which are equivalent to

$$\begin{aligned} y &\leq x + 1/2 \\ y &\leq -x + 1/2 \\ -y &\leq 1/2; \end{aligned}$$

see Figure 5.4.

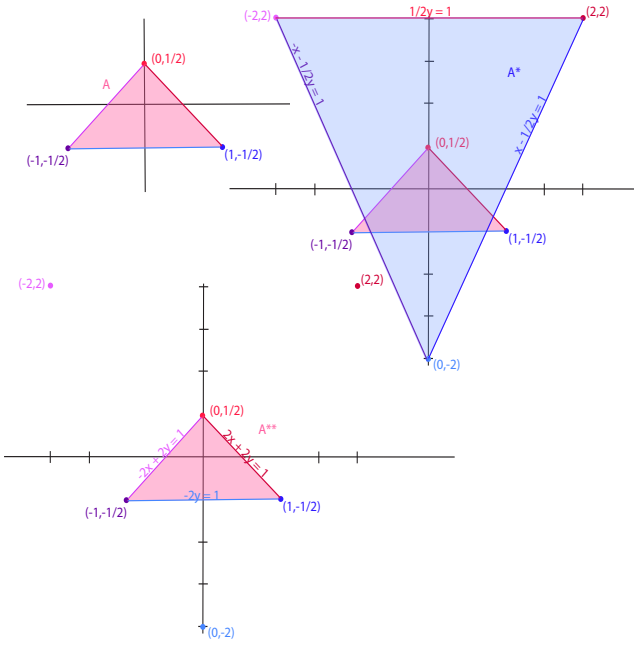


Figure 5.4: The triangle of Example 5.1 written as both a $\mathcal{V}$-polyhedron and an $\mathcal{H}$-polyhedron.

Example 5.2. Let A be the $\mathcal{V}$-polyhedron (a square) in $\mathbb{A}^2$ defined by set $Y = \{(-1/2, 0), (0, -1/2), (1/2, 0), (0, 1/2)\}$. By Proposition 5.14, the polar dual A^* is the $\mathcal{H}$-polytope, another square, cut out by the inequalities:

$$\begin{aligned} (1/2)y &\leq 1 \\ -(1/2)x &\leq 1 \\ -(1/2)y &\leq 1 \\ (1/2)x &\leq 1. \end{aligned}$$

This is also the $\mathcal{V}$-polytope whose vertices are $(-2, 2), (-2, -2), (2, -2), (2, 2)$. By Proposition 5.6, $A = A^{**}$ is $\mathcal{H}$-polyhedron cut out by the inequalities

$$\begin{aligned} -2x + 2y &\leq 1 \\ -2x - 2y &\leq 1 \\ 2x - 2y &\leq 1 \\ 2x + 2y &\leq 1, \end{aligned}$$

which are equivalent to

$$\begin{aligned} y &\leq x + 1/2 \\ y &\geq -x - 1/2 \\ y &\geq x - 1/2 \\ y &\leq -x + 1/2; \end{aligned}$$

see Figure 5.5.

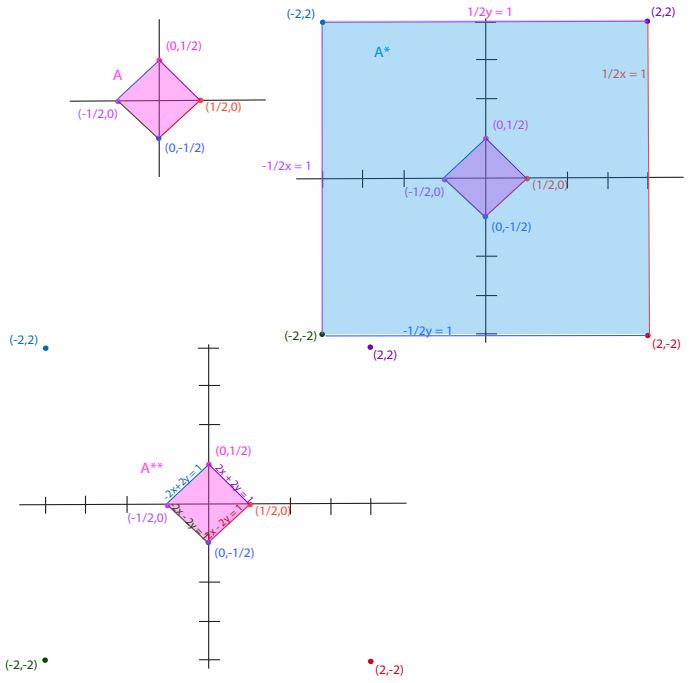


Figure 5.5: The diamond of Example 5.2 written as both a $\mathcal{V}$-polyhedron and an $\mathcal{H}$-polyhedron.

Example 5.3. Let A be the $\mathcal{V}$-polyhedron in $\mathbb{A}^2$ defined by the set $Y = \{(0, -1)\}$ consisting of a single point and the set of vectors $V = \{(-1, 1), (1, 1)\}$. This polyhedron is a cone with apex $(0, -1)$. By Proposition 5.14, the polar dual A^* is the $\mathcal{H}$-polytope, a triangle, cut out by the inequalities:

$$\begin{aligned} -x + y &\leq 0 \\ x + y &\leq 0 \\ -y &\leq 1. \end{aligned}$$

This is also the $\mathcal{V}$-polytope whose vertices are $(0,0), (-1,-1)$, and $(1,-1)$. By Proposition 5.6, $A = A^{**}$ is $\mathcal{H}$-polyhedron (a cone) cut out by the inequalities

$$\begin{aligned} -x - y &\leq 1 \\ x - y &\leq 1; \end{aligned}$$

see Figure 5.6. Note that there is no line associated with $(0,0)$ since this point belongs to the boundary of the triangle.

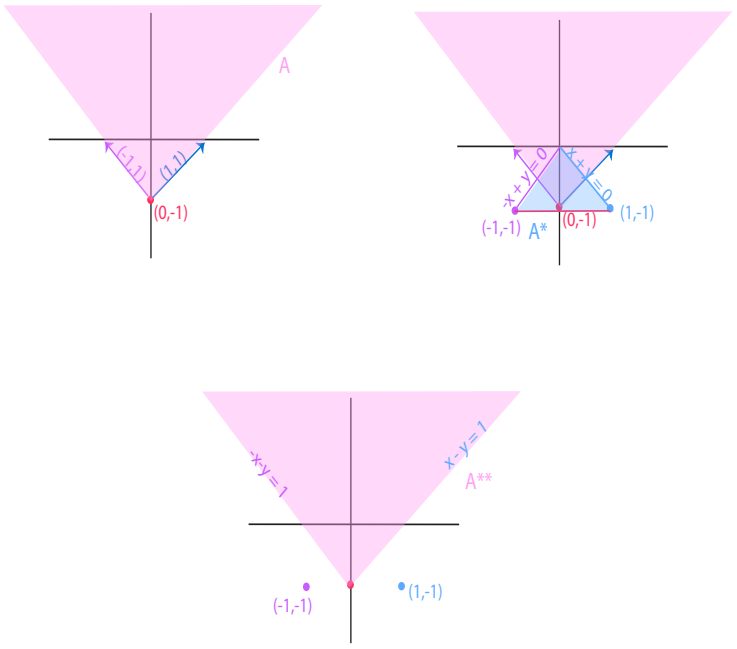


Figure 5.6: The triangular cone of Example 5.3 written as both a $\mathcal{V}$-polyhedron and an $\mathcal{H}$-polyhedron.

Example 5.4. Let A be the $\mathcal{V}$-polyhedron in $\mathbb{A}^2$ defined by set $Y = \{(-1,-1),(1,-1)\}$ and the set of vectors $V = \{(-1,1),(1,1)\}$. By Proposition 5.14, the polar dual A^* is the $\mathcal{H}$-polytope, a convex polygon, cut out by the inequalities:

$$\begin{aligned} -x + y &\leq 0 \\ x + y &\leq 0 \\ -x - y &\leq 1 \\ x - y &\leq 1. \end{aligned}$$

This is also the $\mathcal{V}$-polytope whose vertices are $(0,0), (-1/2,-1/2), (0,-1), (1/2,-1/2)$. By Proposition 5.6, $A = A^{**}$ is $\mathcal{H}$-polyhedron cut out by the inequalities

$$\begin{aligned} -(1/2)x - (1/2)y &\leq 1 \\ -y &\leq 1 \\ (1/2)x - (1/2)y &\leq 1, \end{aligned}$$

which are equivalent to

$$\begin{aligned} y &\geq -x - 2 \\ y &\geq -1 \\ y &\geq x - 2; \end{aligned}$$

see Figure 5.7.

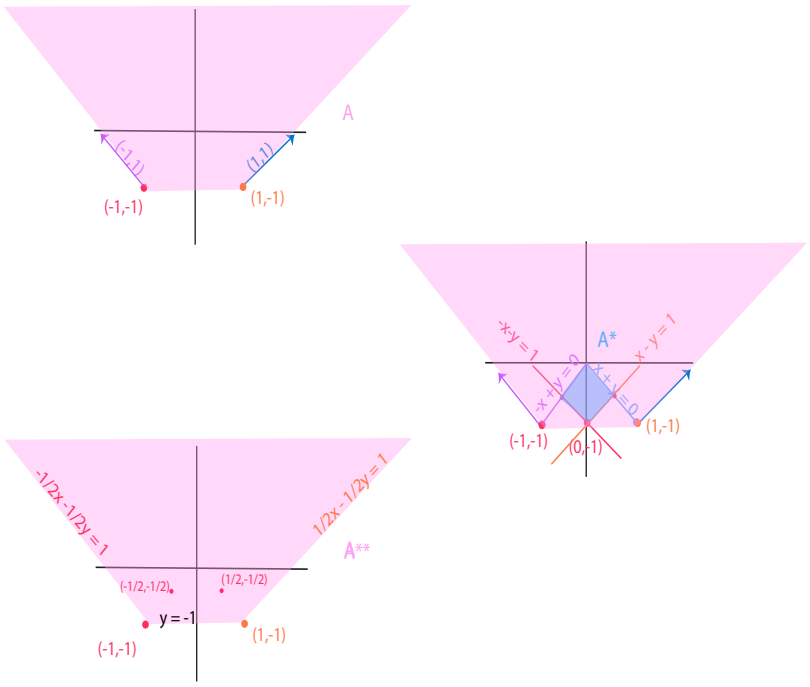


Figure 5.7: The trough of Example 5.4 written as both a $\mathcal{V}$-polyhedron and an $\mathcal{H}$-polyhedron.

Example 5.5. Let A be the $\mathcal{V}$-polyhedron in $\mathbb{A}^2$ defined by sets $Y = \{(0,1), (-1,0), (1,0)\}$ and $V = \{(0,-1)\}$. By Proposition 5.14, the polar dual A^* is the $\mathcal{H}$-polytope, a square, cut

out by the inequalities:

$$\begin{aligned} y &\leq 1 \\ -x &\leq 1 \\ x &\leq 1 \\ -y &\leq 0. \end{aligned}$$

This is also the $\mathcal{V}$-polytope (square) whose vertices are $(-1,1), (-1,0), (1,0), (1,1)$. By Proposition 5.6, $A = A^{**}$ is $\mathcal{H}$-polyhedron cut out by the inequalities

$$\begin{aligned} -x + y &\leq 1 \\ -x &\leq 1 \\ x &\leq 1 \\ x + y &\leq 1; \end{aligned}$$

see Figure 5.8.

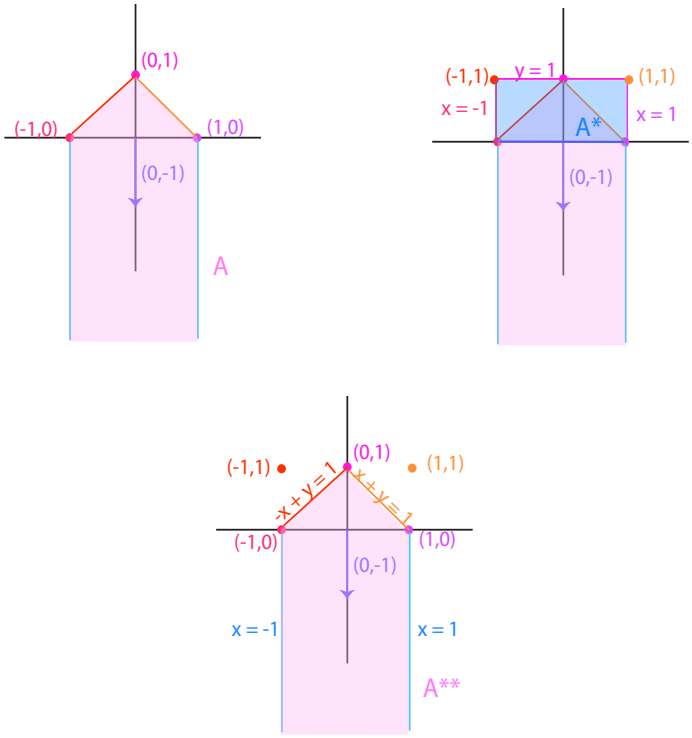


Figure 5.8: The triangular peaked of Example 5.5 written as both a $\mathcal{V}$-polyhedron and an $\mathcal{H}$-polyhedron.

In all the previous examples, the step that is not automatic is to find the vertices of an $\mathcal{H}$-polytope from the inequalities defining this $\mathcal{H}$-polytope. For small examples in dimension 2 this is easy, but in general this is an expansive process. A method to do this algorithmically is Fourier–Motzkin elimination; see Proposition 5.21.

Here are now some example illustrating Proposition 5.18.

Example 5.6. Let A be the $\mathcal{H}$-polyhedron in $\mathbb{A}^2$ defined by the inequalities

$$\begin{aligned} -x - y &\leq 1 \\ x - y &\leq 1. \end{aligned}$$

This is the cone arising in Example 5.3. By Proposition 5.6, the polar dual A^* a $\mathcal{V}$-polytope, a triangle, the convex hull of the points $(0,0), (-1,-1)$, and $(1,-1)$. In Example 5.1, we computed the equations of the triangle $(0,1/2), (-1,-1/2)$, and $(1,-1/2)$ obtained by translating the above triangle by $(0,1/2)$, namely

$$\begin{aligned} y &\leq x + 1/2 \\ y &\leq -x + 1/2 \\ -y &\leq 1/2, \end{aligned}$$

so the triangle $(0,0), (-1,-1)$, and $(1,-1)$ is also the $\mathcal{H}$-polyhedron (triangle) defined by the inequalities

$$\begin{aligned} -x + y &\leq 0 \\ x + y &\leq 0 \\ -y &\leq 1, \end{aligned}$$

and by Proposition 5.15, $A = A^{**}$ is the $\mathcal{V}$-polyhedron given by the set $Y = \{(0,-1)\}$ consisting of a single point and the set of vectors $V = \{(-1,1),(1,1)\}$; see Figure 5.9.

Example 5.7. Let A be the $\mathcal{H}$-polyhedron in $\mathbb{A}^2$ defined by the inequalities

$$\begin{aligned} -(1/2)x - (1/2)y &\leq 1 \\ -y &\leq 1 \\ (1/2)x - (1/2)y &\leq 1. \end{aligned}$$

This is the $\mathcal{H}$-polyhedron of Example 5.4. By Proposition 5.6, the polar dual A^* a $\mathcal{V}$-polytope, the square $(0,0), (-1/2,-1/2), (0,-1), (1/2,-1/2)$. In Example 5.2, we computed the equations of the square $(-1/2,0)$, $(0,-1/2), (1/2,0), (0,1/2)$ obtained by translating the above square by $(0,1/2)$, namely

$$\begin{aligned} y &\leq x + 1/2 \\ y &\geq -x - 1/2 \\ y &\geq x - 1/2 \\ y &\leq -x + 1/2, \end{aligned}$$

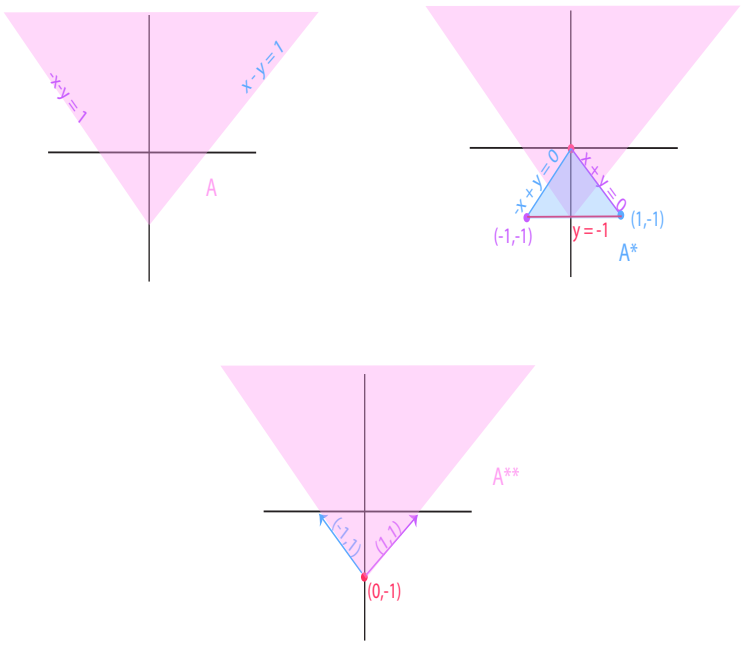


Figure 5.9: The triangular cone of Example 5.6 written as first an $\mathcal{H}$-polyhedron and then $\mathcal{V}$-polyhedron.

so the square $(0,0),(-1/2,-1/2),(0,-1),(1/2,-1/2)$ is also the $\mathcal{H}$-polyhedron (square) defined by the inequalities

$$\begin{aligned} -x+y &\leq 0\\ -x-y &\leq 1\\ x-y &\leq 1\\ x+y &\leq 0. \end{aligned}$$

By Proposition 5.15, $A = A^{**}$ is the $\mathcal{V}$-polyhedron given by the set $Y = \{(-1,-1),(1,-1)\}$ and the set of vectors $V = \{(-1,1),(1,1)\}$; see Figure 5.10.

Example 5.8. Let A be the $\mathcal{H}$-polyhedron in $\mathbb{A}^2$ defined by the inequalities

$$\begin{aligned} -x-y &\leq 1\\ -y &\leq 0\\ x-y &\leq 1. \end{aligned}$$

By Proposition 5.6, the polar dual A^* is the $\mathcal{V}$-polyhedron given by the set of points $(0,0),(-1,-1)$, and $(1,-1)$ and the vector $(0,-1)$. In example 5.5, we computed the inequalities of the $\mathcal{V}$-polyedron given by $Y = \{(0,1),(-1,0),(1,0)\}$ and $V = \{(0,-1)\}$, obtained

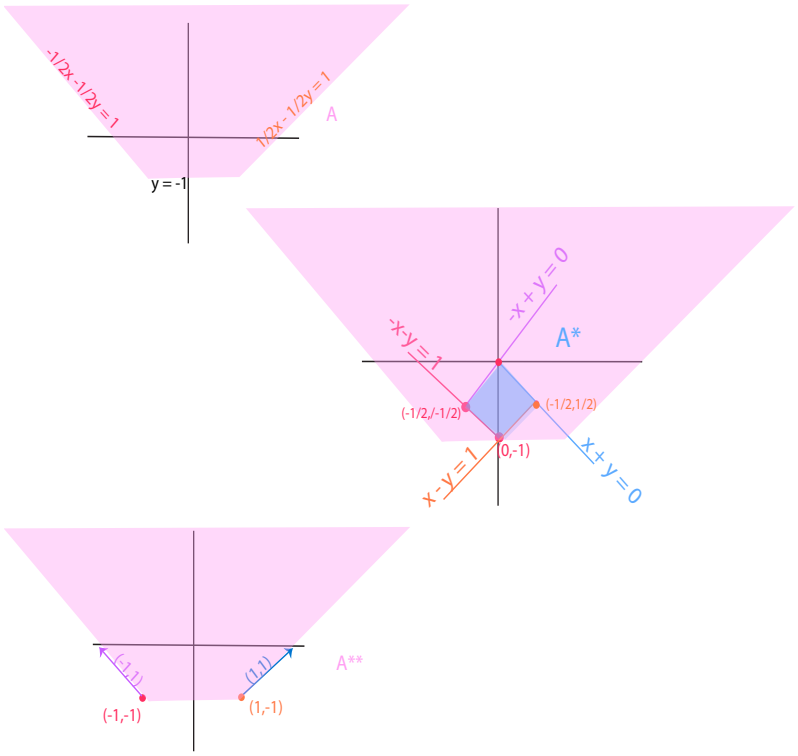


Figure 5.10: The trough of Example 5.4 written as first an $\mathcal{H}$-polyhedron and then $\mathcal{V}$-polyhedron.

by translating the above $\mathcal{V}$-polyhedron by $(0, 1)$, namely

$$\begin{aligned} -x + y &\leq 1 \\ -x &\leq 1 \\ x &\leq 1 \\ x + y &\leq 1, \end{aligned}$$

so the $\mathcal{V}$-polyhedron given by the set of points $(0, 0), (-1, -1)$, and $(1, -1)$ and the vector $(0, -1)$ is defined by the inequalities

$$\begin{aligned} -x + y &\leq 0 \\ -x &\leq 1 \\ x &\leq 1 \\ x + y &\leq 0. \end{aligned}$$

By Proposition 5.15, $A = A^{**}$ is the $\mathcal{V}$-polyhedron given by the sets $Y = \{(-1, 0), (1, 0)\}$ and $V = \{(-1, 1), (1, 1)\}$; see Figure 5.11.

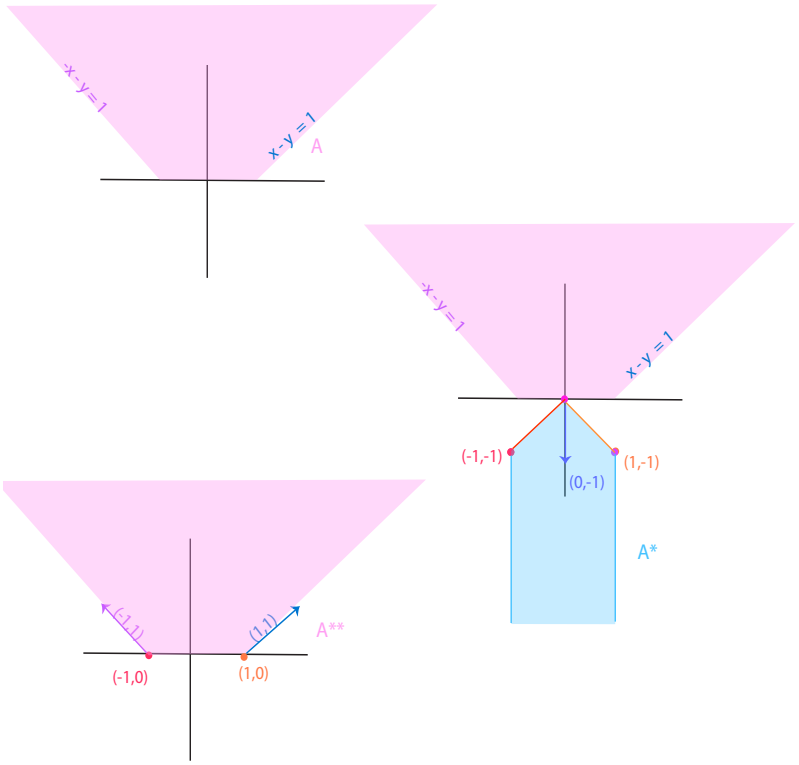


Figure 5.11: The trough of Example 5.8 written as first an $\mathcal{H}$-polyhedron and then $\mathcal{V}$-polyhedron.

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*.

5.5 Fourier–Motzkin Elimination and the Polyhedron-Cone Correspondence

The problem with the converse of Proposition 5.17 when A is unbounded (*i.e.*, not compact) is that Krein and Milman's Theorem does not apply. We need to take into account "points at infinity" corresponding to certain vectors. The trick we used in Proposition 5.17 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 Milman 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. If S is given by m polynomial inequalities of the form

$$P_i(x_1, \ldots, x_d) \leq b_i,$$

where $P_i(x_1, \ldots, x_d)$ is a polynomial of total degree n_i, this amounts to forming the new homogeneous inequalities

$$x_{d+1}^{n_i} P_i\left(\frac{x_1}{x_{d+1}}, \ldots, \frac{x_d}{x_{d+1}}\right) - b_i x_{d+1}^{n_i} \leq 0$$

together with $x_{d+1} \geq 0$. In particular, if the P_i's are linear forms (which means that $n_i = 1$), then our inequalities are of the form

$$a_i \cdot x \leq b_i,$$

where $a_i \in \mathbb{R}^d$ is some vector, and the homogenized inequalities are

$$a_i \cdot x - b_i x_{d+1} \leq 0.$$

It will be convenient to formalize the process of putting a copy of a subset $S \subseteq \mathbb{E}^d$ in the hyperplane $H_{d+1} \subseteq \mathbb{E}^{d+1}$ of equation $x_{d+1} = 1$ as follows: For every point $a \in \mathbb{E}^d$, let

$$\widehat{a} = \begin{pmatrix} a \\ 1 \end{pmatrix} \in \mathbb{E}^{d+1},$$

and let $\widehat{S} = \{\widehat{a} \mid a \in S\}$. Obviously, the map $S \mapsto \widehat{S}$ is a bijection between the subsets of $\mathbb{E}^d$ and the subsets of H_{d+1}. We will use this bijection to identify S and $\widehat{S}$ and use the simpler notation S, unless confusion arises. Let's apply this to polyhedra.

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

$$\begin{aligned} C(P) &= \left\{ \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} \in \mathbb{R}^{d+1} \mid Ax \leq x_{d+1} b,\ x_{d+1} \geq 0 \right\} \\ &= \left\{ \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} \in \mathbb{R}^{d+1} \mid Ax - x_{d+1} b \leq 0,\ -x_{d+1} \leq 0 \right\}. \end{aligned}$$

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

$$\widehat{P} = C(P) \cap H_{d+1}.$$

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 $H_{d+1} \approx \mathbb{E}^d$.

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

$$\widehat{P} = C(P) \cap H_{d+1},$$

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 $w_{i\,d+1} \geq 0$ for every $w_i \in W$, we prove next that $P = C \cap H_{d+1}$ is a $\mathcal{V}$-polyhedron.

Proposition 5.20. *(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 $\widehat{P} = C(P) \cap H_{d+1}$, 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 $H_{d+1} \approx \mathbb{E}^d$.

(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

$$\widehat{P} = C(P) \cap H_{d+1},$$

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 $H_{d+1} \approx \mathbb{E}^d$.

In both (1) and (2), $\widehat{P} = \{\widehat{p} \mid p \in P\}$, with

$$\widehat{p} = \begin{pmatrix} p \\ 1 \end{pmatrix} \in \mathbb{E}^{d+1}.$$

Proof. We already proved everything except the last part of the proposition. Let

$$\widehat{Y} = \left\{ \frac{w_i}{w_{i\,d+1}} \;\middle|\; w_i \in W,\ w_{i\,d+1} > 0 \right\}$$

and

$$\widehat{V} = \{w_j \in W \mid w_{j\,d+1} = 0\}.$$

We claim that

$$P = C \cap H_{d+1} = \mathrm{conv}(\widehat{Y}) + \mathrm{cone}(\widehat{V}),$$

and thus, P is a $\mathcal{V}$-polyhedron.

Recall that any element $z \in C$ can be written as

$$z = \sum_{k=1}^{s} \mu_k w_k, \quad \mu_k \geq 0.$$

Thus, we have

$$\begin{aligned}
z &= \sum_{k=1}^{s} \mu_k w_k \quad \mu_k \geq 0 \\
&= \sum_{w_{i\,d+1}>0} \mu_i w_i + \sum_{w_{j\,d+1}=0} \mu_j w_j \quad \mu_i, \mu_j \geq 0 \\
&= \sum_{w_{i\,d+1}>0} w_{i\,d+1}\mu_i \frac{w_i}{w_{i\,d+1}} + \sum_{w_{j\,d+1}=0} \mu_j w_j, \quad \mu_i, \mu_j \geq 0 \\
&= \sum_{w_{i\,d+1}>0} \lambda_i \frac{w_i}{w_{i\,d+1}} + \sum_{w_{j\,d+1}=0} \mu_j w_j, \quad \lambda_i, \mu_j \geq 0.
\end{aligned}$$

Now, $z \in C \cap H_{d+1}$ iff $z_{d+1} = 1$ iff $\sum_{i=1}^{p} \lambda_i = 1$ (where p is the number of elements in $\widehat{Y}$), since the $(d+1)^{\text{th}}$ coordinate of $\frac{w_i}{w_{i\,d+1}}$ is equal to 1, and the above shows that

$$P = C \cap H_{d+1} = \mathrm{conv}(\widehat{Y}) + \mathrm{cone}(\widehat{V}),$$

as claimed. □

By Proposition 5.20, 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 5.20 shows that $\widehat{P} = C(P) \cap H_{d+1}$ is a $\mathcal{V}$-polyhedron and so P 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 5.20 shows 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 5.17.

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 an affine subspace in $\mathbb{E}^{d+m}$.

We claim that $C_0(A)$ is a $\mathcal{V}$-cone. This follows by observing that for every $\binom{x}{w}$ satisfying $Ax \leq w$, we can write

$$\begin{pmatrix} x \\ w \end{pmatrix} = \sum_{i=1}^{d} |x_i|(\operatorname{sign}(x_i)) \begin{pmatrix} e_i \\ Ae_i \end{pmatrix} + \sum_{j=1}^{m} (w_j - (Ax)_j) \begin{pmatrix} 0 \\ e_j \end{pmatrix},$$

and then

$$C_0(A) = \operatorname{cone}\left(\left\{ \pm \begin{pmatrix} e_i \\ Ae_i \end{pmatrix} \mid 1 \leq i \leq d \right\} \cup \left\{ \begin{pmatrix} 0 \\ e_j \end{pmatrix} \mid 1 \leq j \leq m \right\} \right).$$

Since $C = C_0(A) \cap H_w$ is now the intersection of a $\mathcal{V}$-cone with an affine subspace, 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 5.21. *(Fourier–Motzkin Elimination) Say $C = \mathrm{cone}(Y) \subseteq \mathbb{E}^d$ 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 $C \cap H_k = \mathrm{cone}(Y^{/k})$, 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 $\mathbb{R}^d$.

Proof. The only nontrivial direction is to prove that $C \cap H_k \subseteq \mathrm{cone}(Y^{/k})$. For this, consider any $v = \sum_{i=1}^{d} t_i y_i \in C \cap H_k$, with $t_i \geq 0$ and $v_k = 0$. Such a v can be written

$$v = \sum_{i|y_{ik}=0} t_i y_i + \sum_{i|y_{ik}>0} t_i y_i + \sum_{j|y_{jk}<0} t_j y_j$$

and as $v_k = 0$, we have

$$\sum_{i|y_{ik}>0} t_i y_{ik} + \sum_{j|y_{jk}<0} t_j y_{jk} = 0.$$

If $t_i y_{ik} = 0$ for $i = 1, \ldots, d$, we are done. Otherwise, we can write

$$\Lambda = \sum_{i|y_{ik}>0} t_i y_{ik} = \sum_{j|y_{jk}<0} -t_j y_{jk} > 0.$$

Then,

$$\begin{aligned} v &= \sum_{i|y_{ik}=0} t_i y_i + \frac{1}{\Lambda} \sum_{i|y_{ik}>0} \left(\sum_{j|y_{jk}<0} -t_j y_{jk} \right) t_i y_i + \frac{1}{\Lambda} \sum_{j|y_{jk}<0} \left(\sum_{i|y_{ik}>0} t_i y_{ik} \right) t_j y_j \\ &= \sum_{i|y_{ik}=0} t_i y_i + \sum_{\substack{i|y_{ik}>0 \\ j|y_{jk}<0}} \frac{t_i t_j}{\Lambda} \left(y_{ik}y_j - y_{jk}y_i \right). \end{aligned}$$

Since the k^{th} coordinate of $y_{ik}y_j - y_{jk}y_i$ is 0, the above shows that any $v \in C \cap H_k$ can be written as a positive combination of vectors in $Y^{/k}$. □

As discussed above, Proposition 5.21 implies (again!)

Corollary 5.22. *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. This method is interesting in its own right so we discuss it briefly.

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. Here again, it is possible that $P = \mathbb{E}^d$, but that's fine since $\mathbb{E}^d$ has been declared to be an $\mathcal{H}$-polyhedron.

In view of Proposition 5.20 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 5.21. We state the result without proof and refer the interested reader to Ziegler [69], Section 1.2–1.3, for full details.

Proposition 5.23. *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 $\mathrm{elim}_k(C) = \{x \in \mathbb{R}^d \mid (\exists t \in \mathbb{R})(x + te_k \in P)\} = \{z - te_k \mid z \in P,\ t \in \mathbb{R}\} = P(A^{/k}, 0)$ 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. Therefore, it is important to find ways of detecting redundancies and there are various methods for doing so. Again, the interested reader should consult Ziegler [69], Chapter 1.

There is yet another way of proving that an $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron without using Fourier–Motzkin elimination that was inspired to us by the proof of Theorem 1.5 in Chapter II of Ewald [26]. As we already observed, Krein and Milman's theorem does not apply if our polyhedron is unbounded. Actually, the full power of Krein and Milman's theorem is not needed to show that an $\mathcal{H}$-polytope is a $\mathcal{V}$-polytope.

The crucial point is that if P is an $\mathcal{H}$-polytope with nonempty interior, then *every* line ℓ through any point a in the interior of P intersects P in a line segment. This is because P is compact and ℓ is closed, so $P \cap \ell$ is a compact subset of a line thus, a closed interval $[b, c]$ with $b < a < c$, as a is in the interior of P. Therefore, we can use induction on the dimension of P to show that every point in P is a convex combination of vertices of the facets of P.

Now, if P is unbounded and cut out by at least two half-spaces (so, P is not a half-space), then we claim that for every point a in the interior of P, there is *some* line through a that

intersects two facets of P. This is because if we pick the origin in the interior of P, we may assume that P is given by an irredundant intersection, $P = \bigcap_{i=1}^{t}(H_i)_-$, and for any point a in the interior of P, there is a line ℓ through P *in general position w.r.t.* P, which means that ℓ is not parallel to any of the hyperplanes H_i and intersects all of them in distinct points (see Definition 11.2). Fortunately, lines in general position always exist, as shown in Proposition 11.3. Using this fact, we can prove the following result:

Proposition 5.24. *Let $P \subseteq \mathbb{E}^d$ be an $\mathcal{H}$-polyhedron $P = \bigcap_{i=1}^{t}(H_i)_-$ (an irredundant decomposition), with nonempty interior. If $t = 1$, that is, $P = (H_1)_-$ is a half-space, then*

$$P = a + \mathrm{cone}(u_1, \ldots, u_{d-1}, -u_1, \ldots, -u_{d-1}, u_d),$$

where a is any point in H_1, the vectors $u_1, \ldots, u_{d-1}$ form a basis of the direction of H_1, and u_d is normal to (the direction of) H_1. (When $d = 1$, P is the half-line, $P = \{a + tu_1 \mid t \geq 0\}$.) If $t \geq 2$, then every point $a \in P$ can be written as a convex combination $a = (1-\alpha)b + \alpha c$ $(0 \leq \alpha \leq 1)$, where b and c belong to two distinct facets F and G of P, and where

$$F = \mathrm{conv}(Y_F) + \mathrm{cone}(V_F) \quad \textit{and} \quad G = \mathrm{conv}(Y_G) + \mathrm{cone}(V_G),$$

for some finite sets of points Y_F and Y_G and some finite sets of vectors V_F and V_G. (Note: $\alpha = 0$ or $\alpha = 1$ is allowed.) Consequently, every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron.

Proof. We proceed by induction on the dimension d of P. If $d = 1$, then P is either a closed interval $[b, c]$, or a half-line $\{a + tu \mid t \geq 0\}$, where $u \neq 0$. In either case, the proposition is clear.

For the induction step, assume $d > 1$. Since every facet F of P has dimension $d-1$, the induction hypothesis holds for F, that is, there exist a finite set of points Y_F, and a finite set of vectors V_F, so that

$$F = \mathrm{conv}(Y_F) + \mathrm{cone}(V_F).$$

Thus, every point on the boundary of P is of the desired form. Next, pick any point a in the interior of P. Then, from our previous discussion, there is a line ℓ through a in general position w.r.t. P. Consequently, the intersection points $z_i = \ell \cap H_i$ of the line ℓ with the hyperplanes supporting the facets of P exist and are all distinct. If we give ℓ an orientation, the z_i's can be sorted. Since ℓ contains a which is in the interior of P, any point on ℓ to the left of z_1 must be outside P, and similarly any point of ℓ to the right of the rightmost z_i, say z_N, must be outside P, since otherwise there would be a hyperplane cutting ℓ before z_1 or a hyperplane cutting ℓ after z_N. Since P is closed and convex and ℓ is closed, $P \cap \ell$ is a closed an convex subset of ℓ. But this subset is bounded since all points outside $[z_1, z_N]$ are outside P. It follows that $P \cap \ell$ is a closed interval $[b, c]$ with $b, c \in P$, so there is a unique k such that a lies between $b = z_k$ and $c = z_{k+1}$. But then, $b \in F_k = F$ and $c \in F_{k+1} = G$, where F and G the facets of P supported by H_k and H_{k+1}, and $a = (1-\alpha)b + \alpha c$, a convex combination.

Consequently, every point in P is a mixed convex + positive combination of finitely many points associated with the facets of P and finitely many vectors associated with the

directions of the supporting hyperplanes of the facets P. Conversely, it is easy to see that any such mixed combination must belong to P and therefore, P is a $\mathcal{V}$-polyhedron. □

We conclude this section with a version of Farkas Lemma for polyhedral sets.

Lemma 5.25. *(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 $c^\top z < \alpha$.

Proof. We use Farkas Lemma, Version II (Lemma 4.16). Observe that (a) is equivalent to the problem: Find $(u, t) \in \mathbb{R}^{p+q}$, so that

$$\begin{pmatrix} u \\ t \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \mathbb{I} & \mathbb{O} \\ Y & V \end{pmatrix} \begin{pmatrix} u \\ t \end{pmatrix} = \begin{pmatrix} 1 \\ z \end{pmatrix},$$

which is exactly Farkas II(a). Now, the second alternative of Farkas II says that there is no solution as above if there is some $(-\alpha, c) \in \mathbb{R}^{d+1}$ so that

$$(-\alpha, c^\top) \begin{pmatrix} 1 \\ z \end{pmatrix} < 0 \quad \text{and} \quad (-\alpha, c^\top) \begin{pmatrix} \mathbb{I} & 0 \\ Y & V \end{pmatrix} \geq (\mathbb{O}, \mathbb{O}).$$

These are equivalent to

$$-\alpha + c^\top z < 0, \quad -\alpha\mathbb{I} + c^\top Y \geq \mathbb{O}, \quad c^\top V \geq \mathbb{O},$$

namely, $c^\top z < \alpha$, $c^\top Y \geq \alpha\mathbb{I}$ and $c^\top V \geq \mathbb{O}$, which are indeed the conditions of Farkas IV(b), in matrix form. □

Observe that Farkas IV can be viewed as a separation criterion for polyhedral sets. This version subsumes Farkas I and Farkas II.

5.6 Lineality Space and Recession Cone

Given a $\mathcal{V}$-polyhedron $P = \text{conv}(Y) + \text{cone}(V)$, the lines or the rays contained in P play an important role which suggests the following definition.

Definition 5.5. Given a convex set $A \subseteq \mathbb{E}^d$, the *lineality space* of A is defined by

$$\text{lineal}(A) = \{v \in \mathbb{R}^d \mid x + tv \in A \text{ for all } x \in A, \text{ and all } t \in \mathbb{R}\},$$

and the *recession cone* of A is defined by

$$\text{rec}(A) = \{v \in \mathbb{R}^d \mid x + tv \in A \text{ for all } x \in A, \text{ and all } t \geq 0, \text{ where } t \in \mathbb{R}\}.$$

It is immediate from these definitions that lineal(A) is a linear subspace of $\mathbb{R}^d$ and that rec(A) is a convex cone in $\mathbb{R}^d$ containing 0.

If we pick a subspace U of $\mathbb{E}^d$ such that U and lineal(A) form a direct sum $\mathbb{E}^d = \text{lineal}(A) \oplus U$, for example, the othogonal complement of lineal(A), we can decompose A as

$$A = \text{lineal}(A) + (U \cap A),$$

with lineal$(U \cap A) = (0)$. A convex set A such that lineal$(A) = (0)$ is said to be *pointed*.

If P is an $\mathcal{H}$-polyhedron of the form $P = P(A, d)$, then it is immediate by definition that

$$\text{lineal}(P) = \text{Ker}\, A = \{x \in \mathbb{R}^d \mid Ax = 0\}.$$

Regarding recession cones, we have the following proposition.

Proposition 5.26. *Let $P \subseteq \mathbb{E}^d$ be a convex set.*

(1) If P is an $\mathcal{H}$-polyhedron of the form $P = P(A, b)$, then the recession cone rec(P) *is given by*

$$\text{rec}(P) = P(A, \mathbf{0}).$$

(2) If P is a $\mathcal{V}$-polyhedron of the form $P = \text{conv}(Y) + \text{cone}(V)$, then the recession cone rec(P) *is given by*

$$\text{rec}(P) = \text{cone}(V).$$

Proof. The only part whose proof is nontrivial is the inclusion rec$(P) \subseteq$ cone(V). We prove the contrapositive, if $v \notin \text{cone}(V)$ then $v \notin \text{rec}(P)$, using Farkas Lemma Version IV.

By Farkas Lemma Version IV (Proposition 5.25) with $Y = \emptyset$, if $v \notin \text{cone}(V)$, then there is some $c \in \mathbb{R}^d$ and some $\alpha \in \mathbb{R}$ such that $c^\top v_j \geq 0$ for $j = 1, \ldots, q$, $0 \geq \alpha$, and $c^\top v < \alpha$. This implies that $-c^\top v_j \leq 0$ for $j = 1, \ldots, q$, and $-c^\top v > 0$. Let $d = -c$.

For any $x \in P = \text{conv}(Y) + \text{cone}(V)$, we have

$$x = \sum_{i=1}^{p} \lambda_i y_i + \sum_{j=1}^{q} \mu_j v_j,$$

with $\lambda_i, \mu_j \geq 0$ and $\sum_{i=1}^{p} \lambda_i = 1$. Since $d^\top v_j \leq 0$ for $j = 1, \ldots, q$ and $0 \leq \lambda_i \leq 1$, we get

$$d^\top x = \sum_{i=1}^{p} \lambda_i d^\top y_i + \sum_{j=1}^{q} \mu_j d^\top v_j \leq \sum_{i=1}^{p} \lambda_i d^\top y_i \leq \max_{1 \leq i \leq p} d^\top y_i = K,$$

where K is a constant that depends only on d and Y. However,

$$d^\top (x + tv) = d^\top x + t d^\top v,$$

and since $d^\top v > 0$, we see that $d^\top (x + tv)$ tends to $+\infty$ when t tends to $+\infty$, which implies that $x + tv \notin P$, and which means that $v \notin \text{rec}(P)$. □

Proposition 5.26 shows that if P is a $\mathcal{V}$-polyhedron then $v \in \mathrm{rec}(P)$ iff there is *some* $x \in P$ such that $x + tv \in P$ for all $t \geq 0$, which is much cheaper to check that the condition of Definition 5.5 which requires checking that $x + tv \in P$ for all $t \geq 0$ and for *all* $x \in P$.

Notes. The treatment of polytopes and polyhedra given in this chapter is based on the following texts (in alphabetic order): Barvinok [4], Berger [8], Ewald [26], Grunbaum [36], and Ziegler [69]. The terminology $\mathcal{V}$-polyhedron, $\mathcal{V}$-polytope, $\mathcal{H}$-polyhedron, $\mathcal{H}$-polytope, is borrowed from Ziegler.

The proof of Theorem 5.7 (Weyl-Minkowski) using Krein and Millman's theorem and polar duality is taken from Berger [8] (Chapter 12, Proposition 12.1.5). Rather different proofs of the fact that every $\mathcal{V}$-polytope is an $\mathcal{H}$-polytope are given in Grunbaum [36] (Chapter 3, Theorem 3.1.1), and Ewald [26] (Chapter II, Theorems 1.4 and 1.5).

We believe that the proof of Proposition 5.17 showing that every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron using Proposition 5.14, Krein and Millman's theorem, and double dualization, is new. However, this proof is not that original in the sense that it uses the double dualization trick already found in Berger, and the crucial observation that the polar dual A^* of a $\mathcal{V}$-polyhedron A with respect to a center in the interior of A is a bounded $\mathcal{H}$-polyhedron, that is, an $\mathcal{H}$-polytope. We also believe that the proof of Proposition 5.18 showing that every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron is new (it uses a quadruple polar dualization!).

The equivalence of $\mathcal{V}$-polyhedra is and $\mathcal{H}$-polyhedra is also treated (using different techniques) in some books on convex optimization, among which we recommend Bertsekas [9].

Except for Proposition 5.24, which we believe is new, the results of Section 5.5 on Fourier–Motzkin elimination and the polyhedron-cone correspondence are taken from Ziegler [69]. Similarly, the material of Section 5.6 is taken from Ziegler [69].

Chapter 6

Linear Programming

6.1 What is Linear Programming?

What is *linear programming*? At first glance, one might think that this is some style of computer programming. After all, there is imperative programming, functional programming, object-oriented programming, *etc.* The term linear programming is somewhat misleading, because it really refers to a method for *planning* with linear constraints, or more accurately, an *optimization method* where both the objective function and the constraints are linear.[1]

Linear programming was created in the late 1940's, one of the key players being George Dantzing, who invented the simplex algorithm. Kantorovitch also did some pioneering work on linear programming as early as 1939. The term *linear programming* has a military connotation because in the early 1950's it was used as a synonym for plans or schedules for training troops, logistical supply, resource allocation, *etc.* Unfortunately the term linear programming is well established and we are stuck with it.

Interestingly, even though originally most applications of linear programming were in the field of economics and industrial engineering, linear programming has become an important tool in theoretical computer science and in the theory of algorithms. Indeed, linear programming is often an effective tool for designing approximation algorithms to solve hard problems (typically NP-hard problems). Linear programming is also the "baby version" of convex programming, a very effective methodology which has received much attention in recent years.

Our goal is to present the mathematical underpinnings of linear programming, in particular the existence of an optimal solution if a linear program is feasible and bounded, and the duality theorem in linear programming, one of the deepest results in this field. The duality theorem in linear programming also has significant algorithmic implications but we do not discuss this here. We present the simplex algorithm, the dual simplex algorithm, and the primal dual algorithm. We also describe the tableau formalism for running the simplex

[1] Again, we witness another unfortunate abuse of terminology; the constraints are in fact *affine*.

algorithm and its variants. A particularly nice feature of the tableau formalism is that the update of a tableau can be performed using elementary row operations identical to the operations used during the reduction of a matrix to row reduced echelon form (rref). What differs is the criterion for the choice of the pivot.

However, we do not discuss other methods such as the ellipsoid method or interior points methods. For these more algorithmic issues, we refer the reader to standard texts on linear programming. In our opinion, one of the clearest (and among the most concise!) is Matousek and Gardner [42]; Chvatal [18] and Schrijver [53] are classics. Papadimitriou and Steiglitz [47] offers a very crisp presentation in the broader context of combinatorial optimization, and Bertsimas and Tsitsiklis [10] and Vanderbei [66] are very complete.

Linear programming has to do with maximizing a linear cost function $c_1x_1 + \cdots + c_nx_n$ with respect to m "linear" inequalities of the form

$$a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i.$$

These constraints can be put together into an $m \times n$ matrix $A = (a_{ij})$, and written more concisely as

$$Ax \leq b.$$

For technical reasons that will appear clearer later on, it is often preferable to add the nonnegativity constaints $x_i \geq 0$ for $i = 1, \ldots, n$. We write $x \geq 0$. It is easy to show that every linear program is equivalent to another one satisfying the constraints $x \geq 0$, at the expense of adding new variables that are also constrained to be nonnegative. Let $\mathcal{P}(A, b)$ be the set of *feasible solutions* of our linear program given by

$$\mathcal{P}(A, b) = \{x \in \mathbb{R}^n \mid Ax \leq b,\ x \geq 0\}.$$

Then there are two basic questions:

(1) Is $\mathcal{P}(A, b)$ nonempty, that is, does our linear program have a chance to have a solution?

(2) Does the objective function $c_1x_1 + \cdots + c_nx_n$ have a maximum value on $\mathcal{P}(A, b)$?

The answer to both questions can be **no**. But if $\mathcal{P}(A, b)$ is nonempty and if the objective function is bounded above (on $\mathcal{P}(A, b)$), then it can be shown that the maximum of $c_1x_1 + \cdots + c_nx_n$ is achieved by some $x \in \mathcal{P}(A, b)$. Such a solution is called an *optimal solution*. Perhaps surprisingly, this result is not so easy to prove (unless one has the simplex method at his disposal). We will prove this result in full detail (see Proposition 7.1).

The reason why linear constraints are so important is that the domain of potential optimal solutions $\mathcal{P}(A, b)$ is *convex*. In fact, $\mathcal{P}(A, b)$ is a convex polyhedron which is the intersection of half-spaces cut out by affine hyperplanes. The objective function being linear is convex, and this is also a crucial fact. Thus, we are led to study convex sets, in particular those that arise from solutions of inequalities defined by affine forms, but also convex cones.

We give a brief introduction to these topics. As a reward, we provide several criteria for testing whether a system of inequalities

$$Ax \leq b,\ x \geq 0$$

has a solution or not in terms of versions of the *Farkas lemma* (see Proposition 4.14 and Proposition 4.16). Then we give a complete proof of the strong duality theorem for linear programming (see Theorem 9.7). We also discuss the complementary slackness conditions and show that they can be exploited to design an algorithm for solving a linear program that uses both the primal problem and its dual. This algorithm known as the *primal dual algorithm*, although not used much nowadays, has been the source of inspiration for a whole class of approximation algorithms also known as primal dual algorithms.

We hope that this chapter and the next three will be a motivation for learning more about linear programming, convex optimization, but also convex geometry. The "bible" in convex optimization is Boyd and Vandenberghe [14], and one of the best sources for convex geometry is Ziegler [69]. This is a rather advanced text, so the reader may want to begin with Gallier [31].

6.2 Notational Preliminaries

We view $\mathbb{R}^n$ as consisting of *column vectors* ($n \times 1$ matrices). As usual, row vectors represent *linear forms*, that is linear maps $\varphi\colon \mathbb{R}^n \to \mathbb{R}$, in the sense that the row vector y (a $1 \times n$ matrix) represents the linear form φ if $\varphi(x) = yx$ for all $x \in \mathbb{R}^n$. We denote the space of linear forms (row vectors) by $(\mathbb{R}^n)^*$.

Definition 6.1. An *affine form* $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ is defined by some linear form $c \in (\mathbb{R}^n)^*$ and some scalar $\beta \in \mathbb{R}$ so that

$$\varphi(x) = cx + \beta \quad \text{for all } x \in \mathbb{R}^n.$$

If $c \neq 0$, the affine form φ specified by (c, β) defines the *affine hyperplane* (for short *hyperplane*) $H(\varphi)$ given by

$$H(\varphi) = \{x \in \mathbb{R}^n \mid \varphi(x) = 0\} = \{x \in \mathbb{R}^n \mid cx + \beta = 0\},$$

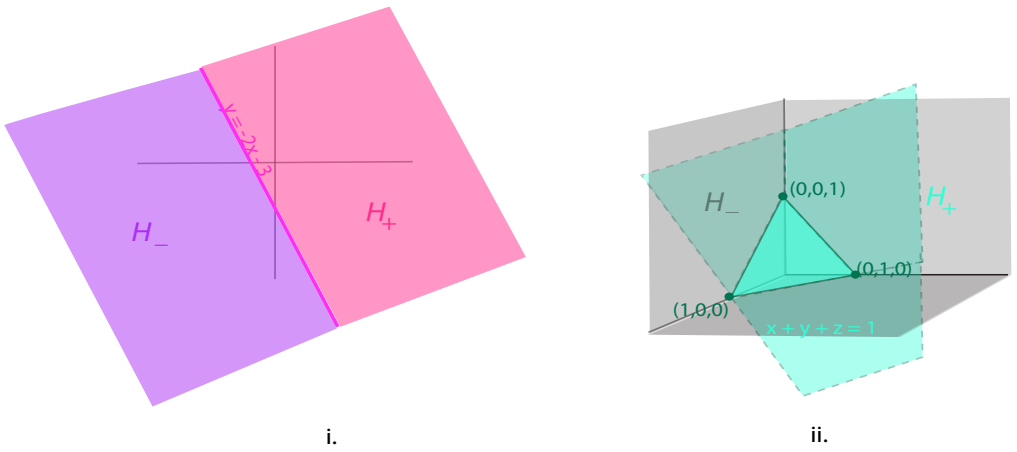
and the two *(closed) half-spaces*

$$H_+(\varphi) = \{x \in \mathbb{R}^n \mid \varphi(x) \geq 0\} = \{x \in \mathbb{R}^n \mid cx + \beta \geq 0\},$$
$$H_-(\varphi) = \{x \in \mathbb{R}^n \mid \varphi(x) \leq 0\} = \{x \in \mathbb{R}^n \mid cx + \beta \leq 0\}.$$

When $\beta = 0$, we call H a *linear hyperplane*.

Both $H_+(\varphi)$ and $H_-(\varphi)$ are convex and $H = H_+(\varphi) \cap H_-(\varphi)$.

For example, $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ with $\varphi(x, y) = 2x + y + 3$ is an affine form defining the line given by the equation $y = -2x - 3$. Another example of an affine form is $\varphi\colon \mathbb{R}^3 \to \mathbb{R}$ with $\varphi(x, y, z) = x + y + z - 1$; this affine form defines the plane given by the equation $x + y + z = 1$, which is the plane through the points $(0, 0, 1), (0, 1, 0)$, and $(1, 0, 0)$. Both of these hyperplanes are illustrated in Figure 6.1.



i. ii.

Figure 6.1: Figure i. illustrates the hyperplane $H(\varphi)$ for $\varphi(x, y) = 2x + y + 3$, while Figure ii. illustrates the hyperplane $H(\varphi)$ for $\varphi(x, y, z) = x + y + z - 1$.

Definition 6.2. For any two vector $x, y \in \mathbb{R}^n$ with $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ we write $x \leq y$ iff $x_i \leq y_i$ for $i = 1, \ldots, n$, and $x \geq y$ iff $y \leq x$. In particular $x \geq 0$ iff $x_i \geq 0$ for $i = 1, \ldots, n$.

6.3 Summary

The main concepts and results of this chapter are listed below:

- Affine form.
- Affine hyperplane, half-spaces.
-

Chapter 7

Linear Programs

In this chapter we introduce linear programs and the basic notions relating to this concept. We define the $\mathcal{H}$-polyhedron $\mathcal{P}(A, b)$ of feasible solutions. Then we define bounded and unbounded linear programs and the notion of optimal solution. We define slack variables and the important notion of *linear program in standard form*.

We show that if a linear program in standard form has a feasible solution and is bounded above, then it has an optimal solution. This is not an obvious result and the proof relies on the fact that a polyhedral cone is closed (this result was shown in the previous chapter).

Next we show that in order to find optimal solutions it suffices to consider solutions of a special form called *basic feasible solutions*. We prove that if a linear program in standard form has a feasible solution and is bounded above, then some basic feasible solution is an optimal solution (Theorem 7.4).

Geometrically, a basic feasible solution corresponds to a *vertex*. In Theorem 7.6 we prove that a basic feasible solution of a linear program in standard form is a vertex of the polyhedron $\mathcal{P}(A, b)$. Finally, we prove that if a linear program in standard form has some feasible solution, then it has a basic feasible solution (see Theorem 7.7). This fact allows the simplex algorithm described in the next chapter to get started.

7.1 Linear Programs, Feasible Solutions, Optimal Solutions

The purpose of linear programming is to solve the following type of optimization problem.

Definition 7.1. *A Linear Program (P)* is the following kind of optimization problem:

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \\ &\quad a_1 x \leq b_1 \\ &\quad \cdots \\ &\quad a_m x \leq b_m \\ &\quad x \geq 0, \end{aligned}$$

where $x \in \mathbb{R}^n$, $c, a_1, \ldots, a_m \in (\mathbb{R}^n)^*$, $b_1, \ldots, b_m \in \mathbb{R}$.

The linear form c defines the *objective function* $x \mapsto cx$ of the Linear Program (P) (from $\mathbb{R}^n$ to $\mathbb{R}$), and the inequalities $a_i x \leq b_i$ and $x_j \geq 0$ are called the *constraints* of the Linear Program (P).

If we define the $m \times n$ matrix

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$$

whose rows are the row vectors $a_1, \ldots, a_m$ and b as the column vector

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},$$

the m inequality constraints $a_i x \leq b_i$ can be written in matrix form as

$$Ax \leq b.$$

Thus the Linear Program (P) can also be stated as the Linear Program (P):

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0. \end{aligned}$$

We should note that in many applications, the natural primal optimization problem is actually the *minimization* of some objective function $cx = c_1 x_1 + \cdots + c_n x_n$, rather its maximization. For example, many of the optimization problems considered in Papadimitriou and Steiglitz [47] are minimization problems.

Of course, minimizing cx is equivalent to maximizing $-cx$, so our presentation covers minimization too.

Here is an explicit example of a linear program of Type (P):

Example 7.1.

$$\begin{aligned}
&\text{maximize} && x_1 + x_2 \\
&\text{subject to} && \\
& && x_2 - x_1 \leq 1 \\
& && x_1 + 6x_2 \leq 15 \\
& && 4x_1 - x_2 \leq 10 \\
& && x_1 \geq 0,\ x_2 \geq 0,
\end{aligned}$$

and in matrix form

$$\begin{aligned}
&\text{maximize} && \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&\text{subject to} && \\
& && \begin{pmatrix} -1 & 1 \\ 1 & 6 \\ 4 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leq \begin{pmatrix} 1 \\ 15 \\ 10 \end{pmatrix} \\
& && x_1 \geq 0,\ x_2 \geq 0.
\end{aligned}$$

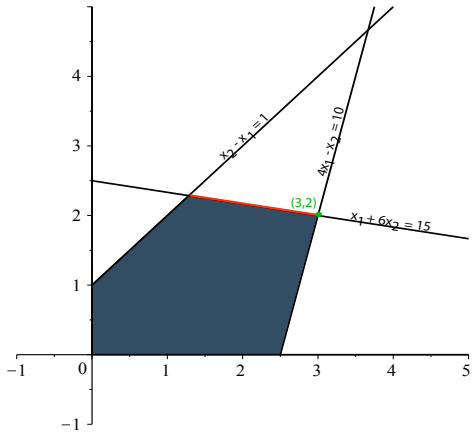


Figure 7.1: The $\mathcal{H}$-polyhedron associated with Example 7.1. The green point $(3, 2)$ is the unique optimal solution.

It turns out that $x_1 = 3, x_2 = 2$ yields the maximum of the objective function $x_1 + x_2$, which is 5. This is illustrated in Figure 7.1. Observe that the set of points that satisfy the above constraints is a convex region cut out by half planes determined by the lines of

equations

$$\begin{aligned} x_2 - x_1 &= 1 \\ x_1 + 6x_2 &= 15 \\ 4x_1 - x_2 &= 10 \\ x_1 &= 0 \\ x_2 &= 0. \end{aligned}$$

In general, each constraint $a_i x \leq b_i$ corresponds to the affine form φ_i given by $\varphi_i(x) = a_i x - b_i$ and defines the half-space $H_-(\varphi_i)$, and each inequality $x_j \geq 0$ defines the half-space $H_+(x_j)$. The intersection of these half-spaces is the set of solutions of all these constraints. It is a (possibly empty) $\mathcal{H}$-polyhedron denoted $\mathcal{P}(A, b)$.

Definition 7.2. If $\mathcal{P}(A, b) = \emptyset$, we say that the Linear Program (P) has *no feasible solution*, and otherwise any $x \in \mathcal{P}(A, b)$ is called a *feasible solution* of (P).

The linear program shown in Example 7.2 obtained by reversing the direction of the inequalities $x_2 - x_1 \leq 1$ and $4x_1 - x_2 \leq 10$ in the linear program of Example 7.1 has no feasible solution; see Figure 7.2.

Example 7.2.

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} \\ & && x_1 - x_2 \leq -1 \\ & && x_1 + 6x_2 \leq 15 \\ & && x_2 - 4x_1 \leq -10 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Assume $\mathcal{P}(A, b) \neq \emptyset$, so that the Linear Program (P) has a feasible solution. In this case, consider the image $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ of $\mathcal{P}(A, b)$ under the objective function $x \mapsto cx$.

Definition 7.3. If the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is unbounded above, then we say that the Linear Program (P) is *unbounded*.

The linear program shown in Example 7.3 obtained from the linear program of Example 7.1 by deleting the constraints $4x_1 - x_2 \leq 10$ and $x_1 + 6x_2 \leq 15$ is unbounded.

Example 7.3.

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} \\ & && x_2 - x_1 \leq 1 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

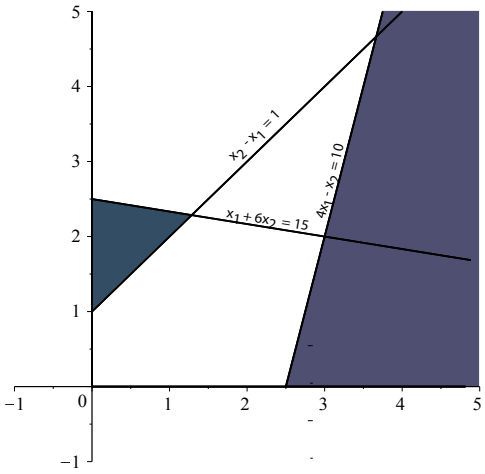


Figure 7.2: There is no $\mathcal{H}$-polyhedron associated with Example 7.2 since the blue and purple regions do not overlap.

Otherwise, we will prove shortly that if μ is the least upper bound of the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, then there is some $p \in \mathcal{P}(A, b)$ such that

$$cp = \mu,$$

that is, the objective function $x \mapsto cx$ has a maximum value μ on $\mathcal{P}(A, b)$ which is achieved by some $p \in \mathcal{P}(A, b)$.

Definition 7.4. If the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is nonempty and bounded above, any point $p \in \mathcal{P}(A, b)$ such that $cp = \max\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is called an *optimal solution* (or *optimum*) of (P). Optimal solutions are often denoted by an upper $*$; for example, p^*.

The linear program of Example 7.1 has a unique optimal solution $(3, 2)$, but observe that the linear program of Example 7.4 in which the objective function is $(1/6)x_1 + x_2$ has infinitely many optimal solutions; the maximum of the objective function is $15/6$ which occurs along the points of orange boundary line in Figure 7.1.

Example 7.4.

$$\begin{aligned} &\text{maximize} && \frac{1}{6}x_1 + x_2 \\ &\text{subject to} \\ & && x_2 - x_1 \leq 1 \\ & && x_1 + 6x_2 \leq 15 \\ & && 4x_1 - x_2 \leq 10 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

The proof that if the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is nonempty and bounded above, then there is an optimal solution $p \in \mathcal{P}(A, b)$, is not as trivial as it might seem. It relies on the fact that a polyhedral cone is closed, a fact that was shown in Section 4.1.

We also use a trick that makes the proof simpler, which is that a Linear Program (P) with inequality constraints $Ax \leq b$

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

is equivalent to the Linear Program (P_2) with equality constraints

$$\begin{aligned} &\text{maximize} && \widehat{c}\,\widehat{x} \\ &\text{subject to} && \widehat{A}\widehat{x} = b \text{ and } \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ is an $m \times (n+m)$ matrix, $\widehat{c}$ is a linear form in $(\mathbb{R}^{n+m})^*$, and $\widehat{x} \in \mathbb{R}^{n+m}$, given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{c} = \begin{pmatrix} c & 0_m^\top \end{pmatrix}, \quad \text{and} \quad \widehat{x} = \begin{pmatrix} x \\ z \end{pmatrix},$$

with $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$.

Indeed, $\widehat{A}\widehat{x} = b$ and $\widehat{x} \geq 0$ iff

$$Ax + z = b, \quad x \geq 0,\ z \geq 0,$$

iff

$$Ax \leq b, \quad x \geq 0,$$

and $\widehat{c}\,\widehat{x} = cx$.

Definition 7.5. The variables z are called *slack variables*, and a linear program of the form (P_2) is called a linear program in *standard form*.

The result of converting the linear program of Example 7.4 to standard form is the program shown in Example 7.5.

Example 7.5.

$$\begin{aligned} &\text{maximize} && \frac{1}{6}x_1 + x_2 \\ &\text{subject to} && \\ & && x_2 - x_1 + z_1 = 1 \\ & && x_1 + 6x_2 + z_2 = 15 \\ & && 4x_1 - x_2 + z_3 = 10 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ z_1 \geq 0,\ z_2 \geq 0,\ z_3 \geq 0. \end{aligned}$$

We can now prove that if a linear program has a feasible solution and is bounded, then it has an optimal solution.

Proposition 7.1. *Let (P_2) be a linear program in standard form, with equality constraint $Ax = b$. If $\mathcal{P}(A, b)$ is nonempty and bounded above, and if μ is the least upper bound of the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, then there is some $p \in \mathcal{P}(A, b)$ such that*

$$cp = \mu,$$

that is, the objective function $x \mapsto cx$ has a maximum value μ on $\mathcal{P}(A, b)$ which is achieved by some optimum solution $p \in \mathcal{P}(A, b)$.

Proof. Since $\mu = \sup\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, there is a sequence $(x^{(k)})_{k\geq 0}$ of vectors $x^{(k)} \in \mathcal{P}(A, b)$ such that $\lim_{k\mapsto\infty} cx^{(k)} = \mu$. In particular, if we write $x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})$ we have $x_j^{(k)} \geq 0$ for $j = 1, \ldots, n$ and for all $k \geq 0$. Let $\widetilde{A}$ be the $(m + 1) \times n$ matrix

$$\widetilde{A} = \begin{pmatrix} c \\ A \end{pmatrix},$$

and consider the sequence $(\widetilde{A}x^{(k)})_{k\geq 0}$ of vectors $\widetilde{A}x^{(k)} \in \mathbb{R}^{m+1}$. We have

$$\widetilde{A}x^{(k)} = \begin{pmatrix} c \\ A \end{pmatrix} x^{(k)} = \begin{pmatrix} cx^{(k)} \\ Ax^{(k)} \end{pmatrix} = \begin{pmatrix} cx^{(k)} \\ b \end{pmatrix},$$

since by hypothesis $x^{(k)} \in \mathcal{P}(A, b)$, and the constraints are $Ax = b$ and $x \geq 0$. Since by hypothesis $\lim_{k\mapsto\infty} cx^{(k)} = \mu$, the sequence $(\widetilde{A}x^{(k)})_{k\geq 0}$ converges to the vector $\begin{pmatrix} \mu \\ b \end{pmatrix}$. Now, observe that each vector $\widetilde{A}x^{(k)}$ can be written as the convex combination

$$\widetilde{A}x^{(k)} = \sum_{j=1}^{n} x_j^{(k)} \widetilde{A}^j,$$

with $x_j^{(k)} \geq 0$ and where $\widetilde{A}^j \in \mathbb{R}^{m+1}$ is the jth column of $\widetilde{A}$. Therefore, $\widetilde{A}x^{(k)}$ belongs to the polyheral cone

$$C = \text{cone}(\widetilde{A}^1, \ldots, \widetilde{A}^n) = \{\widetilde{A}x \mid x \in \mathbb{R}^n,\ x \geq 0\},$$

and since by Proposition 4.13 this cone is closed, $\lim_{k\geq\infty} \widetilde{A}x^{(k)} \in C$, which means that there is some $u \in \mathbb{R}^n$ with $u \geq 0$ such that

$$\begin{pmatrix} \mu \\ b \end{pmatrix} = \lim_{k\geq\infty} \widetilde{A}x^{(k)} = \widetilde{A}u = \begin{pmatrix} cu \\ Au \end{pmatrix},$$

that is, $cu = \mu$ and $Au = b$. Hence, u is an optimal solution of (P_2). □

The next question is, how do we find such an optimal solution? It turns out that for linear programs in standard form where the constraints are of the form $Ax = b$ and $x \geq 0$, there are always optimal solutions of a special type called *basic feasible solutions*.

7.2 Basic Feasible Solutions and Vertices

If the system $Ax = b$ has a solution and if some row of A is a linear combination of other rows, then the corresponding equation is redundant, so we may assume that the rows of A are linearly independent; that is, we may assume that A has rank m, so $m \leq n$.

Definition 7.6. If A is an $m \times n$ matrix, for any nonempty subset K of $\{1, \ldots, n\}$, let A_K be the submatrix of A consisting of the columns of A whose indices belong to K. We denote the jth column of the matrix A by A^j.

Definition 7.7. Given a Linear Program (P_2)

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A has rank m, a vector $x \in \mathbb{R}^n$ is a *basic feasible solution* of (P) if $x \in \mathcal{P}(A, b) \neq \emptyset$, and if there is some subset K of $\{1, \ldots, n\}$ of size m such that

(1) The matrix A_K is invertible (that is, the columns of A_K are linearly independent).

(2) $x_j = 0$ for all $j \notin K$.

The subset K is called a *basis* of x. Every index $k \in K$ is called *basic*, and every index $j \notin K$ is called *nonbasic*. Similarly, the columns A^k corresponding to indices $k \in K$ are called *basic*, and the columns A^j corresponding to indices $j \notin K$ are called *nonbasic*. The variables corresponding to basic indices $k \in K$ are called *basic variables*, and the variables corresponding to indices $j \notin K$ are called *nonbasic*.

For example, the linear program

$$\begin{aligned} &\text{maximize} \quad x_1 + x_2 \\ &\text{subject to} \quad x_1 + x_2 + x_3 = 1 \text{ and } x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0, \qquad (*) \end{aligned}$$

has three basic feasible solutions; the basic feasible solution $K = \{1\}$ corresponds to the point $(1, 0, 0)$; the basic feasible solution $K = \{2\}$ corresponds to the point $(0, 1, 0)$; the basic feasible solution $K = \{3\}$ corresponds to the point $(0, 0, 1)$. Each of these points corresponds to the vertices of the slanted purple triangle illustrated in Figure 7.3. The vertices $(1, 0, 0)$ and $(0, 1, 0)$ optimize the objective function with a value of 1.

We now show that if the Standard Linear Program (P_2) as in Definition 7.7 has some feasible solution and is bounded above, then some basic feasible solution is an optimal solution. We follow Matousek and Gardner [42] (Chapter 4, Section 2, Theorem 4.2.3).

First we obtain a more convenient characterization of a basic feasible solution.

Proposition 7.2. *Given any Standard Linear Program (P_2) where $Ax = b$ and A is an $m \times n$ matrix of rank m, for any feasible solution x, if $J_> = \{j \in \{1, \ldots, n\} \mid x_j > 0\}$, then x is a basic feasible solution iff the columns of the matrix $A_{J_>}$ are linearly independent.*

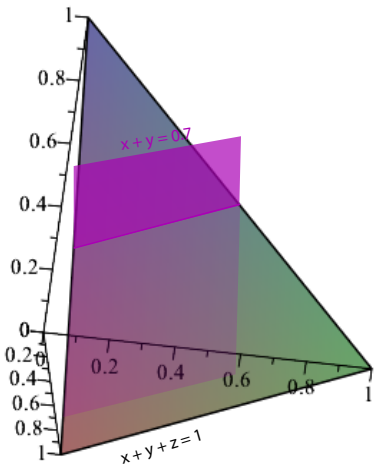


Figure 7.3: The $\mathcal{H}$-polytope associated with Linear Program $(*)$. The objective function (with $x_1 \to x$ and $x_2 \to y$) is represented by vertical planes parallel to the purple plane $x + y = 0.7$, and reaches it maximal value when $x + y = 1$.

Proof. If x is a basic feasible solution, then there is some subset $K \subseteq \{1, \ldots, n\}$ of size m such that the columns of A_K are linearly independent and $x_j = 0$ for all $j \notin K$, so by definition, $J_> \subseteq K$, which implies that the columns of the matrix $A_{J_>}$ are linearly independent.

Conversely, assume that x is a feasible solution such that the columns of the matrix $A_{J_>}$ are linearly independent. If $|J_>| = m$, we are done since we can pick $K = J_>$ and then x is a basic feasible solution. If $|J_>| < m$, we can extend $J_>$ to an m-element subset K by adding $m - |J_>|$ column indices so that the columns of A_K are linearly independent, which is possible since A has rank m. □

Next we prove that if a linear program in standard form has any feasible solution x_0 and is bounded above, then is has some basic feasible solution $\widetilde{x}$ which is as good as x_0, in the sense that $c\widetilde{x} \geq cx_0$.

Proposition 7.3. *Let (P_2) be any standard linear program with objective function cx, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If (P_2) is bounded above and if x_0 is some feasible solution of (P_2), then there is some basic feasible solution $\widetilde{x}$ such that $c\widetilde{x} \geq cx_0$.*

Proof. Among the feasible solutions x such that $cx \geq cx_0$ (x_0 is one of them) pick one with the *maximum* number of coordinates x_j equal to 0, say $\widetilde{x}$. Let

$$K = J_> = \{j \in \{1, \ldots, n\} \mid \widetilde{x}_j > 0\}$$

and let $s = |K|$. We claim that $\widetilde{x}$ is a basic feasible solution, and by construction $c\widetilde{x} \geq cx_0$.

If the columns of A_K are linearly independent, then by Proposition 7.2 we know that $\widetilde{x}$ is a basic feasible solution and we are done.

Otherwise, the columns of A_K are linearly dependent, so there is some nonzero vector $v = (v_1, \ldots, v_s)$ such that $A_K v = 0$. Let $w \in \mathbb{R}^n$ be the vector obtained by extending v by setting $w_j = 0$ for all $j \notin K$. By construction,

$$Aw = A_K v = 0.$$

We will derive a contradiction by exhibiting a feasible solution $x(t_0)$ such that $cx(t_0) \geq cx_0$ with more zero coordinates than $\widetilde{x}$.

For this we claim that we may assume that w satisfies the following two conditions:

(1) $cw \geq 0$.

(2) There is some $j \in K$ such that $w_j < 0$.

If $cw = 0$ and if Condition (2) fails, since $w \neq 0$, we have $w_j > 0$ for some $j \in K$, in which case we can use $-w$, for which $w_j < 0$.

If $cw < 0$, then $c(-w) > 0$, so we may assume that $cw > 0$. If $w_j > 0$ for all $j \in K$, since $\widetilde{x}$ is feasible, $\widetilde{x} \geq 0$, and so $x(t) = \widetilde{x} + tw \geq 0$ for all $t \geq 0$. Furthermore, since $Aw = 0$ and $\widetilde{x}$ is feasible, we have

$$Ax(t) = A\widetilde{x} + tAw = b,$$

and thus $x(t)$ is feasible for all $t \geq 0$. We also have

$$cx(t) = c\widetilde{x} + tcw.$$

Since $cw > 0$, as $t > 0$ goes to infinity the objective function $cx(t)$ also tends to infinity, contradicting the fact that is is bounded above. Therefore, some w satisfying Conditions (1) and (2) above must exist.

We show that there is some $t_0 > 0$ such that $cx(t_0) \geq cx_0$ and $x(t_0) = \widetilde{x} + t_0 w$ is feasible, yet $x(t_0)$ has more zero coordinates than $\widetilde{x}$, a contradiction.

Since $x(t) = \widetilde{x} + tw$, we have

$$x(t)_i = \widetilde{x}_i + tw_i,$$

so if we let $I = \{i \in \{1, \ldots, n\} \mid w_i < 0\} \subseteq K$, which is nonempty since w satisfies Condition (2) above, if we pick

$$t_0 = \min_{i \in I} \left\{ \frac{-\widetilde{x}_i}{w_i} \right\},$$

then $t_0 > 0$, because $w_i < 0$ for all $i \in I$, and by definition of K we have $\widetilde{x}_i > 0$ for all $i \in K$. By the definition of $t_0 > 0$ and since $\widetilde{x} \geq 0$, we have

$$x(t_0)_j = \widetilde{x}_j + t_0 w_j \geq 0 \quad \text{for all } j \in K,$$

so $x(t_0) \geq 0$, and $x(t_0)_i = 0$ for some $i \in I$. Since $Ax(t_0) = b$ (for any t), $x(t_0)$ is a feasible solution,

$$cx(t_0) = c\widetilde{x} + t_0 cw \geq cx_0 + t_0 cw \geq cx_0,$$

and $x(t_0)_i = 0$ for some $i \in I$, we see that $x(t_0)$ has more zero coordinates than $\widetilde{x}$, a contradiction. □

Proposition 7.3 implies the following important result.

Theorem 7.4. *Let (P_2) be any standard linear program with objective function cx, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If (P_2) has some feasible solution and if it is bounded above, then some basic feasible solution $\widetilde{x}$ is an optimal solution of (P_2).*

Proof. By Proposition 7.3, for any feasible solution x there is some basic feasible solution $\widetilde{x}$ such that $cx \leq c\widetilde{x}$. But there are only finitely many basic feasible solutions, so one of them has to yield the maximum of the objective function. □

Geometrically, basic solutions are exactly the vertices of the polyhedron $\mathcal{P}(A, b)$, a notion that we now define.

Definition 7.8. Given an $\mathcal{H}$-polyhedron $\mathcal{P} \subseteq \mathbb{R}^n$, a *vertex* of $\mathcal{P}$ is a point $v \in \mathcal{P}$ with property that there is some nonzero linear form $c \in (\mathbb{R}^n)^*$ and some $\mu \in \mathbb{R}$, such that v is the unique point of $\mathcal{P}$ for which the map $x \mapsto cx$ has the maximum value μ ; that is, $cy < cv = \mu$ for all $y \in \mathcal{P} - \{v\}$. Geometrically, this means that the hyperplane of equation $cy = \mu$ touches $\mathcal{P}$ exactly at v. More generally, a convex subset F of $\mathcal{P}$ is a *k-dimensional face* of $\mathcal{P}$ if F has dimension k and if there is some affine form $\varphi(x) = cx - \mu$ such that $cy = \mu$ for all $y \in F$, and $cy < \mu$ for all $y \in \mathcal{P} - F$. A 1-dimensional face is called an *edge*.

The concept of a vertex is illustrated in Figure 7.4, while the concept of an edge is illustrated in Figure 7.5.

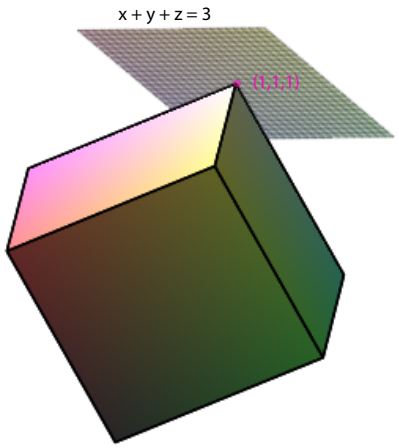


Figure 7.4: The cube centered at the origin with diagonal through $(-1, -1, -1)$ and $(1, 1, 1)$ has eight vertices. The vertex $(1, 1, 1)$ is associated with the linear form $x + y + z = 3$.

Since a k-dimensional face F of $\mathcal{P}$ is equal to the intersection of the hyperplane $H(\varphi)$ of equation $cx = \mu$ with $\mathcal{P}$, it is indeed convex and the notion of dimension makes sense.

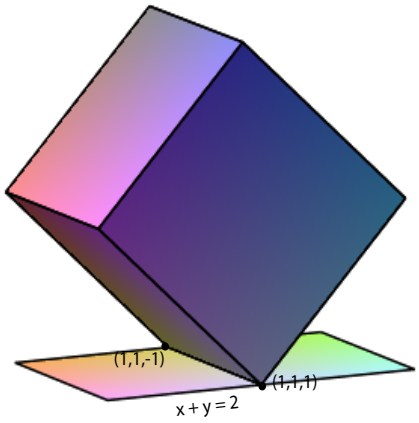


Figure 7.5: The cube centered at the origin with diagonal through $(-1,-1,-1)$ and $(1,1,1)$ has twelve edges. The edge from $(1,1,-1)$ to $(1,1,1)$ is associated with the linear form $x+y=2$.

Observe that a 0-dimensional face of $\mathcal{P}$ is a vertex. If $\mathcal{P}$ has dimension d, then the $(d-1)$-dimensional faces of $\mathcal{P}$ are called its *facets*.

If (P) is a linear program in standard form, then its basic feasible solutions are exactly the vertices of the polyhedron $\mathcal{P}(A,b)$. To prove this fact we need the following simple proposition

Proposition 7.5. *Let $Ax=b$ be a linear system where A is an $m\times n$ matrix of rank m. For any subset $K\subseteq\{1,\ldots,n\}$ of size m, if A_K is invertible, then there is at most one basic feasible solution $x\in\mathbb{R}^n$ with $x_j=0$ for all $j\notin K$ (of course, $x\geq 0$)*

Proof. In order for x to be feasible we must have $Ax=b$. Write $N=\{1,\ldots,n\}-K$, x_K for the vector consisting of the coordinates of x with indices in K, and x_N for the vector consisting of the coordinates of x with indices in N. Then

$$Ax=A_Kx_K+A_Nx_N=b.$$

In order for x to be a basic feasible solution we must have $x_N=0$, so

$$A_Kx_K=b.$$

Since by hypothesis A_K is invertible, $x_K=A_K^{-1}b$ is uniquely determined. If $x_K\geq 0$ then x is a basic feasible solution, otherwise it is not. This proves that there is at most one basic feasible solution $x\in\mathbb{R}^n$ with $x_j=0$ for all $j\notin K$. □

Theorem 7.6. *Let (P) be a linear program in standard form, where $Ax=b$ and A is an $m\times n$ matrix of rank m. For every $v\in\mathcal{P}(A,b)$, the following conditions are equivalent:*

(1) v is a vertex of the Polyhedron $\mathcal{P}(A, b)$.

(2) v is a basic feasible solution of the Linear Program (P).

Proof. First, assume that v is a vertex of $\mathcal{P}(A, b)$, and let $\varphi(x) = cx - \mu$ be a linear form such that $cy < \mu$ for all $y \in \mathcal{P}(A, b)$ and $cv = \mu$. This means that v is the unique point of $\mathcal{P}(A, b)$ for which the objective function $x \mapsto cx$ has the maximum value μ on $\mathcal{P}(A, b)$, so by Theorem 7.4, since this maximum is achieved by some basic feasible solution, by uniqueness v must be a basic feasible solution.

Conversely, suppose v is a basic feasible solution of (P) corresponding to a subset $K \subseteq \{1, \ldots, n\}$ of size m. Let $\widehat{c} \in (\mathbb{R}^n)^*$ be the linear form defined by

$$\widehat{c}_j = \begin{cases} 0 & \text{if } j \in K \\ -1 & \text{if } j \notin K. \end{cases}$$

By construction $\widehat{c}\, v = 0$ and $\widehat{c}\, x \leq 0$ for any $x \geq 0$, hence the function $x \mapsto \widehat{c}\, x$ on $\mathcal{P}(A, b)$ has a maximum at v. Furthermore, $\widehat{c}\, x < 0$ for any $x \geq 0$ such that $x_j > 0$ for some $j \notin K$. However, by Proposition 7.5, the vector v is the only basic feasible solution such that $v_j = 0$ for all $j \notin K$, and therefore v is the only point of $\mathcal{P}(A, b)$ maximizing the function $x \mapsto \widehat{c}\, x$, so it is a vertex. □

In theory, to find an optimal solution we try all $\binom{n}{m}$ possible m-elements subsets K of $\{1, \ldots, n\}$ and solve for the corresponding unique solution x_K of $A_K x = b$. Then we check whether such a solution satisfies $x_K \geq 0$, compute cx_K, and return some feasible x_K for which the objective function is maximum. This is a totally impractical algorithm.

A practical algorithm is the *simplex algorithm*. Basically, the simplex algorithm tries to "climb" in the polyhderon $\mathcal{P}(A, b)$ from vertex to vertex along edges (using basic feasible solutions), trying to maximize the objective function. We present the simplex algorithm in the next chapter. The reader may also consult texts on linear programming. In particular, we recommend Matousek and Gardner [42], Chvatal [18], Papadimitriou and Steiglitz [47], Bertsimas and Tsitsiklis [10], Ciarlet [19], Schrijver [53], and Vanderbei [66].

Observe that Theorem 7.4 asserts that if a Linear Program (P) in standard form (where $Ax = b$ and A is an $m \times n$ matrix of rank m) has some feasible solution and is bounded above, then some basic feasible solution is an optimal solution. By Theorem 7.6, the polyhedron $\mathcal{P}(A, b)$ must have some vertex.

But suppose we only know that $\mathcal{P}(A, b)$ is nonempty; that is, we don't know that the objective function cx is bounded above. Does $\mathcal{P}(A, b)$ have some vertex?

The answer to the above question is *yes*, and this is important because the simplex algorithm needs an initial basic feasible solution to get started. Here we prove that if $\mathcal{P}(A, b)$ is nonempty, then it must contain a vertex. This proof still doesn't constructively yield a vertex, but we will see in the next chapter that the simplex algorithm always finds a vertex if there is one (provided that we use a pivot rule that prevents cycling).

Theorem 7.7. *Let (P) be a linear program in standard form, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If $\mathcal{P}(A, b)$ is nonempty (there is a feasible solution), then $\mathcal{P}(A, b)$ has some vertex; equivalently, (P) has some basic feasible solution.*

Proof. The proof relies on a trick, which is to add slack variables $x_{n+1}, \ldots, x_{n+m}$ and use the new objective function $-(x_{n+1} + \cdots + x_{n+m})$.

If we let $\widehat{A}$ be the $m \times (m+n)$-matrix, and x, $\overline{x}$, and $\widehat{x}$ be the vectors given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n, \quad \overline{x} = \begin{pmatrix} x_{n+1} \\ \vdots \\ x_{n+m} \end{pmatrix} \in \mathbb{R}^m, \quad \widehat{x} = \begin{pmatrix} x \\ \overline{x} \end{pmatrix} \in \mathbb{R}^{n+m},$$

then consider the Linear Program $(\widehat{P})$ in standard form

$$\begin{aligned} &\text{maximize} && -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} && \widehat{A}\,\widehat{x} = b \text{ and } \widehat{x} \geq 0. \end{aligned}$$

Since $x_i \geq 0$ for all i, the objective function $-(x_{n+1} + \cdots + x_{n+m})$ is bounded above by 0. The system $\widehat{A}\,\widehat{x} = b$ is equivalent to the system

$$Ax + \overline{x} = b,$$

so for every feasible solution $u \in \mathcal{P}(A, b)$, since $Au = b$, the vector $(u, 0_m)$ is also a feasible solution of $(\widehat{P})$, in fact an optimal solution since the value of the objective function $-(x_{n+1} + \cdots + x_{n+m})$ for $\overline{x} = 0$ is 0. By Proposition 7.3, the linear program $(\widehat{P})$ has some basic feasible solution (u^*, w^*) for which the value of the objective function is greater than or equal to the value of the objective function for $(u, 0_m)$, and since $(u, 0_m)$ is an optimal solution, (u^*, w^*) is also an optimal solution of $(\widehat{P})$. This implies that $w^* = 0$, since otherwise the objective function $-(x_{n+1} + \cdots + x_{n+m})$ would have a strictly negative value.

Therefore, $(u^*, 0_m)$ is a basic feasible solution of $(\widehat{P})$, and thus the columns corresponding to nonzero components of u^* are linearly independent. Some of the coordinates of u^* could be equal to 0, but since A has rank m we can add columns of A to obtain a basis K associated with u^*, and u^* is indeed a basic feasible solution of (P). □

The definition of a basic feasible solution can be adapted to linear programs where the constraints are of the form $Ax \leq b$, $x \geq 0$; see Matousek and Gardner [42] (Chapter 4, Section 4, Definition 4.4.2).

The most general type of linear program allows constraints of the form $a_i x \geq b_i$ or $a_i x = b_i$ besides constraints of the form $a_i x \leq b_i$. The variables x_i may also take negative values. It is always possible to convert such programs to the type considered in Definition 7.1. We proceed as follows.

Every constraint $a_i x \geq b_i$ is replaced by the constraint $-a_i x \leq -b_i$. Every equality constraint $a_i x = b_i$ is replaced by the two constraints $a_i x \leq b_i$ and $-a_i x \leq -b_i$.

If there are n variables x_i, we create n new variables y_i and n new variables z_i and replace every variable x_i by $y_i - z_i$. We also add the $2n$ constraints $y_i \geq 0$ and $z_i \geq 0$. If the constraints are given by the inequalities $Ax \leq b$, we now have constraints given by

$$\begin{pmatrix} A & -A \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} \leq b, \quad y \geq 0,\ z \geq 0.$$

We replace the objective function cx by $cy - cz$.

Remark: We also showed that we can replace the inequality constraints $Ax \leq b$ by equality constraints $Ax = b$, by adding slack variables constrained to be nonnegative.

7.3 Summary

The main concepts and results of this chapter are listed below:

- Linear program.
- Objective function, constraints.
- Feasible solution.
- Bounded and unbounded linear programs.
- Optimal solution, optimum.
- Slack variables, linear program in standard form.
- Basic feasible solution.
- Basis of a variable.
- Basic, nonbasic index, basic, nonbasic variable.
- Vertex, face, edge, facet.

7.4 Problems

Problem 7.1. Convert the following program to standard form:

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} && \\ & && x_2 - x_1 \leq 1 \\ & && x_1 + 6x_2 \leq 15 \\ & && -4x_1 + x_2 \geq 10. \end{aligned}$$

Problem 7.2. Convert the following program to standard form:

$$\begin{aligned} &\text{maximize} && 3x_1 - 2x_2 \\ &\text{subject to} && \\ & && 2x_1 - x_2 \leq 4 \\ & && x_1 + 3x_2 \geq 5 \\ & && x_2 \geq 0. \end{aligned}$$

Problem 7.3. The notion of basic feasible solution for linear programs where the constraints are of the form $Ax \leq b$, $x \geq 0$ is defined as follows. A basic feasible solution of a (general) linear program with n variables is a feasible solution for which some n linearly independent constraints hold with equality.

Prove that the definition of a basic feasible solution for linear programs in standard form is a special case of the above definition.

Problem 7.4. Consider the linear program

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} && \\ & && x_1 + x_2 \leq 1. \end{aligned}$$

Show that none of the optimal solutions are basic.

Problem 7.5. The *standard n-simplex* is the subset Δ^n of $\mathbb{R}^{n+1}$ given by

$$\Delta^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1 + \cdots + x_{n+1} = 1,\ x_i \geq 0,\ 1 \leq i \leq n+1\}.$$

(1) Prove that Δ^n is convex and that it is the convex hull of the $n+1$ vectors $e_1, \ldots e_{n+1}$, where e_i is the ith canonical unit basis vector, $i = 1, \ldots, n+1$.

(2) Prove that Δ^n is the intersection of $n+1$ half spaces and determine the hyperplanes defining these half-spaces.

Remark: The volume under the standard simplex Δ^n is $1/(n+1)!$.

Problem 7.6. The n-dimensional *cross-polytope* is the subset XP_n of $\mathbb{R}^n$ given by

$$XP_n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid |x_1| + \cdots + |x_n| \leq 1\}.$$

(1) Prove that XP_n is convex and that it is the convex hull of the $2n$ vectors $\pm e_i$, where e_i is the ith canonical unit basis vector, $i = 1, \ldots, n$.

(2) Prove that XP_n is the intersection of 2^n half spaces and determine the hyperplanes defining these half-spaces.

Remark: The volume of XP_n is $2^n/n!$.

Problem 7.7. The n-dimensional *hypercube* is the subset C_n of $\mathbb{R}^n$ given by

$$C_n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid |x_i| \leq 1, 1 \leq i \leq n\}.$$

(1) Prove that C_n is convex and that it is the convex hull of the 2^n vectors $(\pm 1, \ldots, \pm 1)$, $i = 1, \ldots, n$.

(2) Prove that C_n is the intersection of $2n$ half spaces and determine the hyperplanes defining these half-spaces.

Remark: The volume of C_n is 2^n.

Chapter 8

The Simplex Algorithm

8.1 The Idea Behind the Simplex Algorithm

The simplex algorithm, due to Dantzig, applies to a linear program (P) in standard form, where the constraints are given by $Ax = b$ and $x \geq 0$, with A a $m \times n$ matrix of rank m, and with an objective function $x \mapsto cx$. This algorithm either reports that (P) has no feasible solution, or that (P) is unbounded, or yields an optimal solution. Geometrically, the algorithm climbs from vertex to vertex in the polyhedron $\mathcal{P}(A, b)$, trying to improve the value of the objective function. Since vertices correspond to basic feasible solutions, the simplex algorithm actually works with basic feasible solutions.

Recall that a basic feasible solution x is a feasible solution for which there is a subset $K \subseteq \{1, \ldots, n\}$ of size m such that the matrix A_K consisting of the columns of A whose indices belong to K are linearly independent, and that $x_j = 0$ for all $j \notin K$. We also let $J_>(x)$ be the set of indices

$$J_>(x) = \{j \in \{1, \ldots, n\} \mid x_j > 0\},$$

so for a basic feasible solution x associated with K, we have $J_>(x) \subseteq K$. In fact, by Proposition 7.2, a feasible solution x is a basic feasible solution iff the columns of $A_{J_>(x)}$ are linearly independent.

If $J_>(x)$ had cardinality m for all basic feasible solutions x, then the simplex algorithm would make progress at every step, in the sense that it would strictly increase the value of the objective function. Unfortunately, it is possible that $|J_>(x)| < m$ for certain basic feasible solutions, and in this case a step of the simplex algorithm may not increase the value of the objective function. Worse, in rare cases, it is possible that the algorithm enters an infinite loop. This phenomenon called *cycling* can be detected, but in this case the algorithm fails to give a conclusive answer.

Fortunately, there are ways of preventing the simplex algorithm from cycling (for example, Bland's rule discussed later), although proving that these rules work correctly is quite involved.

The potential "bad" behavior of a basic feasible solution is recorded in the following definition.

Definition 8.1. Given a Linear Program (P) in standard form where the constraints are given by $Ax = b$ and $x \geq 0$, with A an $m \times n$ matrix of rank m, a basic feasible solution x is *degenerate* if $|J_>(x)| < m$, otherwise it is *nondegenerate*.

The origin 0_n, if it is a basic feasible solution, is degenerate. For a less trivial example, $x = (0, 0, 0, 2)$ is a degenerate basic feasible solution of the following linear program in which $m = 2$ and $n = 4$.

Example 8.1.

$$\begin{aligned} &\text{maximize} && x_2 \\ &\text{subject to} && \\ & && -x_1 + x_2 + x_3 = 0 \\ & && x_1 + x_4 = 2 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0. \end{aligned}$$

The matrix A and the vector b are given by

$$A = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 2 \end{pmatrix},$$

and if $x = (0, 0, 0, 2)$, then $J_>(x) = \{4\}$. There are two ways of forming a set of two linearly independent columns of A containing the fourth column.

Given a basic feasible solution x associated with a subset K of size m, since the columns of the matrix A_K are linearly independent, by abuse of language we call the columns of A_K a *basis* of x.

If u is a vertex of (P), that is, a basic feasible solution of (P) associated with a basis K (of size m), in "normal mode," the simplex algorithm tries to move along an edge from the vertex u to an adjacent vertex v (with $u, v \in \mathcal{P}(A, b) \subseteq \mathbb{R}^n$) corresponding to a basic feasible solution whose basis is obtained by replacing one of the basic vectors A^k with $k \in K$ by another nonbasic vector A^j for some $j \notin K$, in such a way that the value of the objective function is increased.

Let us demonstrate this process on an example.

Example 8.2. Let (P) be the following linear program in standard form.

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} && \\ & && -x_1 + x_2 + x_3 = 1 \\ & && x_1 + x_4 = 3 \\ & && x_2 + x_5 = 2 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0,\ x_5 \geq 0. \end{aligned}$$

The matrix A and the vector b are given by

$$A = \begin{pmatrix} -1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}.$$

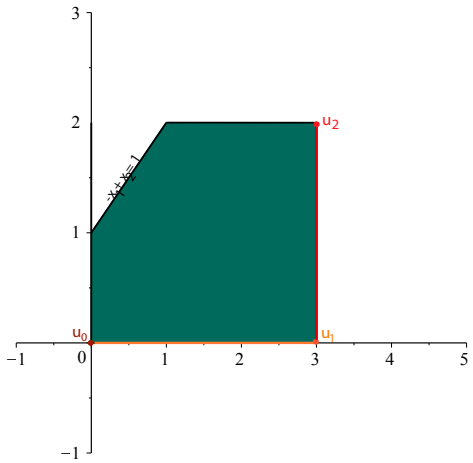


Figure 8.1: The planar $\mathcal{H}$-polyhedron associated with Example 8.2. The initial basic feasible solution is the origin. The simplex algorithm first moves along the horizontal orange line to feasible solution at vertex u_1. It then moves along the vertical red line to obtain the optimal feasible solution u_2.

The vector $u_0 = (0, 0, 1, 3, 2)$ corresponding to the basis $K = \{3, 4, 5\}$ is a basic feasible solution, and the corresponding value of the objective function is $0 + 0 = 0$. Since the columns (A^3, A^4, A^5) corresponding to $K = \{3, 4, 5\}$ are linearly independent we can express A^1 and A^2 as

$$\begin{aligned} A^1 &= -A^3 + A^4 \\ A^2 &= A^3 + A^5. \end{aligned}$$

Since

$$1A^3 + 3A^4 + 2A^5 = Au_0 = b,$$

for any $\theta \in \mathbb{R}$, we have

$$\begin{aligned} b &= 1A^3 + 3A^4 + 2A^5 - \theta A^1 + \theta A^1 \\ &= 1A^3 + 3A^4 + 2A^5 - \theta(-A^3 + A^4) + \theta A^1 \\ &= \theta A^1 + (1 + \theta)A^3 + (3 - \theta)A^4 + 2A^5, \end{aligned}$$

and

$$\begin{aligned} b &= 1A^3 + 3A^4 + 2A^5 - \theta A^2 + \theta A^2 \\ &= 1A^3 + 3A^4 + 2A^5 - \theta(A^3 + A^5) + \theta A^2 \\ &= \theta A^2 + (1-\theta)A^3 + 3A^4 + (2-\theta)A^5. \end{aligned}$$

In the first case, the vector $(\theta, 0, 1+\theta, 3-\theta, 2)$ is a feasible solution iff $0 \leq \theta \leq 3$, and the new value of the objective function is θ.

In the second case, the vector $(0, \theta, 1-\theta, 3, 2-\theta, 1)$ is a feasible solution iff $0 \leq \theta \leq 1$, and the new value of the objective function is also θ.

Consider the first case. It is natural to ask whether we can get another vertex and increase the objective function by setting to zero one of the coordinates of $(\theta, 0, 1+\theta, 3-\theta, 2)$, in this case the fouth one, by picking $\theta = 3$. This yields the feasible solution $(3, 0, 4, 0, 2)$, which corresponds to the basis (A^1, A^3, A^5), and so is indeed a basic feasible solution, with an improved value of the objective function equal to 3. Note that A^4 *left* the basis (A^3, A^4, A^5) and A^1 *entered* the new basis (A^1, A^3, A^5).

We can now express A^2 and A^4 in terms of the basis (A^1, A^3, A^5), which is easy to do since we already have A^1 and A^2 in term of (A^3, A^4, A^5), and A^1 and A^4 are swapped. Such a step is called a *pivoting step*. We obtain

$$\begin{aligned} A^2 &= A^3 + A^5 \\ A^4 &= A^1 + A^3. \end{aligned}$$

Then we repeat the process with $u_1 = (3, 0, 4, 0, 2)$ and the basis (A^1, A^3, A^5). We have

$$\begin{aligned} b &= 3A^1 + 4A^3 + 2A^5 - \theta A^2 + \theta A^2 \\ &= 3A^1 + 4A^3 + 2A^5 - \theta(A^3 + A^5) + \theta A^2 \\ &= 3A^1 + \theta A^2 + (4-\theta)A^3 + (2-\theta)A^5, \end{aligned}$$

and

$$\begin{aligned} b &= 3A^1 + 4A^3 + 2A^5 - \theta A^4 + \theta A^4 \\ &= 3A^1 + 4A^3 + 2A^5 - \theta(A^1 + A^3) + \theta A^4 \\ &= (3-\theta)A^1 + (4-\theta)A^3 + \theta A^4 + 2A^5. \end{aligned}$$

In the first case, the point $(3, \theta, 4-\theta, 0, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the new value of the objective function is $3+\theta$. In the second case, the point $(3-\theta, 0, 4-\theta, \theta, 2)$ is a feasible solution iff $0 \leq \theta \leq 3$, and the new value of the objective function is $3-\theta$. To increase the objective function, we must choose the first case and we pick $\theta = 2$. Then we get the feasible solution $u_2 = (3, 2, 2, 0, 0)$, which corresponds to the basis (A^1, A^2, A^3), and thus is a basic feasible solution. The new value of the objective function is 5.

Next we express A^4 and A^5 in terms of the basis (A^1, A^2, A^3). Again this is easy to do since we just swapped A^5 and A^2 (a pivoting step), and we get

$$A^5 = A^2 - A^3$$
$$A^4 = A^1 + A^3.$$

We repeat the process with $u_2 = (3, 2, 2, 0, 0)$ and the basis (A^1, A^2, A^3). We have

$$\begin{aligned} b &= 3A^1 + 2A^2 + 2A^3 - \theta A^4 + \theta A^4 \\ &= 3A^1 + 2A^2 + 2A^3 - \theta(A^1 + A^3) + \theta A^4 \\ &= (3-\theta)A^1 + 2A^2 + (2-\theta)A^3 + \theta A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 3A^1 + 2A^2 + 2A^3 - \theta A^5 + \theta A^5 \\ &= 3A^1 + 2A^2 + 2A^3 - \theta(A^2 - A^3) + \theta A^5 \\ &= 3A^1 + (2-\theta)A^2 + (2+\theta)A^3 + \theta A^5. \end{aligned}$$

In the first case, the point $(3-\theta, 2, 2-\theta, \theta, 0)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is $5 - \theta$. In the second case, the point $(3, 2-\theta, 2+\theta, 0, \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is also $5 - \theta$. Since we must have $\theta \geq 0$ to have a feasible solution, there is no way to increase the objective function. In this situation, it turns out that we have reached an optimal solution, in our case $u_2 = (3, 2, 2, 0, 0)$, with the maximum of the objective function equal to 5.

We could also have applied the simplex algorithm to the vertex $u_0 = (0, 0, 1, 3, 2)$ and to the vector $(0, \theta, 1-\theta, 3, 2-\theta, 1)$, which is a feasible solution iff $0 \leq \theta \leq 1$, with new value of the objective function θ. By picking $\theta = 1$, we obtain the feasible solution $(0, 1, 0, 3, 1)$, corresponding to the basis (A^2, A^4, A^5), which is indeed a vertex. The new value of the objective function is 1. Then we express A^1 and A^3 in terms the basis (A^2, A^4, A^5) obtaining

$$A^1 = A^4 - A^3$$
$$A^3 = A^2 - A^5,$$

and repeat the process with $(0, 1, 0, 3, 1)$ and the basis (A^2, A^4, A^5). After three more steps we will reach the optimal solution $u_2 = (3, 2, 2, 0, 0)$.

Let us go back to the linear program of Example 8.1 with objective function x_2 and where the matrix A and the vector b are given by

$$A = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

Recall that $u_0 = (0, 0, 0, 2)$ is a degenerate basic feasible solution, and the objective function has the value 0. See Figure 8.2 for a planar picture of the $\mathcal{H}$-polyhedron associated with Example 8.1.

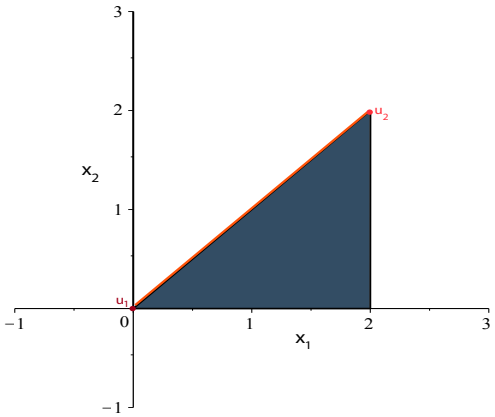


Figure 8.2: The planar $\mathcal{H}$-polyhedron associated with Example 8.1. The initial basic feasible solution is the origin. The simplex algorithm moves along the slanted orange line to the apex of the triangle.

Pick the basis (A^3, A^4). Then we have

$$\begin{aligned} A^1 &= -A^3 + A^4 \\ A^2 &= A^3, \end{aligned}$$

and we get

$$\begin{aligned} b &= 2A^4 - \theta A^1 + \theta A^1 \\ &= 2A^4 - \theta(-A^3 + A^4) + \theta A^1 \\ &= \theta A^1 + \theta A^3 + (2-\theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^4 - \theta A^2 + \theta A^2 \\ &= 2A^4 - \theta A^3 + \theta A^2 \\ &= \theta A^2 - \theta A^3 + 2A^4. \end{aligned}$$

In the first case, the point $(\theta, 0, \theta, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is 0, and in the second case the point $(0, \theta, -\theta, 2)$ is a feasible solution iff $\theta = 0$, and the value of the objective function is θ. However, since we must have $\theta = 0$ in the second case, there is no way to increase the objective function either.

It turns out that in order to make the cases considered by the simplex algorithm as mutually exclusive as possible, since in the second case the coefficient of θ in the value of the objective function is nonzero, namely 1, we should choose the second case. We must

pick $\theta = 0$, but we can swap the vectors A^3 and A^2 (because A^2 is coming in and A^3 has the coefficient $-\theta$, which is the reason why θ must be zero), and we obtain the basic feasible solution $u_1 = (0, 0, 0, 2)$ with the new basis (A^2, A^4). Note that this basic feasible solution corresponds to the same vertex $(0, 0, 0, 2)$ as before, but the basis has changed. The vectors A^1 and A^3 can be expressed in terms of the basis (A^2, A^4) as

$$\begin{aligned} A^1 &= -A^2 + A^4 \\ A^3 &= A^2. \end{aligned}$$

We now repeat the procedure with $u_1 = (0, 0, 0, 2)$ and the basis (A^2, A^4), and we get

$$\begin{aligned} b &= 2A^4 - \theta A^1 + \theta A^1 \\ &= 2A^4 - \theta(-A^2 + A^4) + \theta A^1 \\ &= \theta A^1 + \theta A^2 + (2 - \theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^4 - \theta A^3 + \theta A^3 \\ &= 2A^4 - \theta A^2 + \theta A^3 \\ &= -\theta A^2 + \theta A^3 + 2A^4. \end{aligned}$$

In the first case, the point $(\theta, \theta, 0, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is θ, and in the second case the point $(0, -\theta, \theta, 2)$ is a feasible solution iff $\theta = 0$ and the value of the objective function is θ. In order to increase the objective function we must choose the first case and pick $\theta = 2$. We obtain the feasible solution $u_2 = (2, 2, 0, 0)$ whose corresponding basis is (A^1, A^2) and the value of the objective function is 2.

The vectors A^3 and A^4 are expressed in terms of the basis (A^1, A^2) as

$$\begin{aligned} A^3 &= A^2 \\ A^4 &= A^1 + A^3, \end{aligned}$$

and we repeat the procedure with $u_2 = (2, 2, 0, 0)$ and the basis (A^1, A^2). We get

$$\begin{aligned} b &= 2A^1 + 2A^2 - \theta A^3 + \theta A^3 \\ &= 2A^1 + 2A^2 - \theta A^2 + \theta A^3 \\ &= 2A^1 + (2 - \theta)A^2 + \theta A^3, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^1 + 2A^2 - \theta A^4 + \theta A^4 \\ &= 2A^1 + 2A^2 - \theta(A^1 + A^3) + \theta A^4 \\ &= (2 - \theta)A^1 + 2A^2 - \theta A^3 + \theta A^4. \end{aligned}$$

In the first case, the point $(2, 2-\theta, 0, \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is $2-\theta$, and in the second case, the point $(2-\theta, 2, -\theta, \theta)$ is a feasible solution iff $\theta = 0$ and the value of the objective function is 2. This time there is no way to improve the objective function and we have reached an optimal solution $u_2 = (2, 2, 0, 0)$ with the maximum of the objective function equal to 2.

Let us now consider an example of an unbounded linear program.

Example 8.3. Let (P) be the following linear program in standard form.

$$\begin{aligned} &\text{maximize} && x_1 \\ &\text{subject to} && \\ & && x_1 - x_2 + x_3 = 1 \\ & && -x_1 + x_2 + x_4 = 2 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0. \end{aligned}$$

The matrix A and the vector b are given by

$$A = \begin{pmatrix} 1 & -1 & 1 & 0 \\ -1 & 1 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

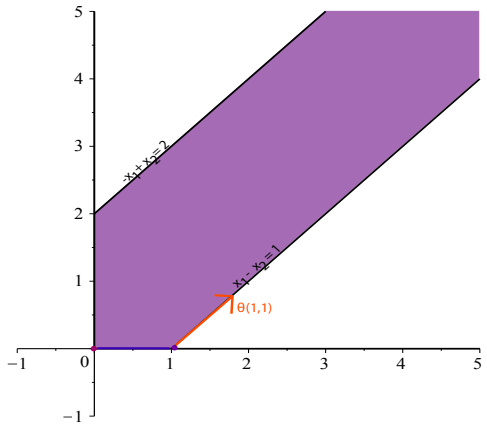


Figure 8.3: The planar $\mathcal{H}$-polyhedron associated with Example 8.3. The initial basic feasible solution is the origin. The simplex algorithm first moves along the horizontal indigo line to basic feasible solution at vertex $(1, 0)$. Any optimal feasible solution occurs by moving along the boundary line parameterized by the orange arrow $\theta(1, 1)$.

The vector $u_0 = (0, 0, 1, 2)$ corresponding to the basis $K = \{3, 4\}$ is a basic feasible solution, and the corresponding value of the objective function is 0. The vectors A^1 and A^2

are expressed in terms of the basis (A^3, A^4) by

$$\begin{aligned} A^1 &= A^3 - A^4 \\ A^2 &= -A^3 + A^4. \end{aligned}$$

Starting with $u_0 = (0, 0, 1, 2)$, we get

$$\begin{aligned} b &= A^3 + 2A^4 - \theta A^1 + \theta A^1 \\ &= A^3 + 2A^4 - \theta(A^3 - A^4) + \theta A^1 \\ &= \theta A^1 + (1 - \theta)A^3 + (2 + \theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= A^3 + 2A^4 - \theta A^2 + \theta A^2 \\ &= A^3 + 2A^4 - \theta(-A^3 + A^4) + \theta A^2 \\ &= \theta A^2 + (1 + \theta)A^3 + (2 - \theta)A^4. \end{aligned}$$

In the first case, the point $(\theta, 0, 1 - \theta, 2 + \theta)$ is a feasible solution iff $0 \leq \theta \leq 1$ and the value of the objective function is θ, and in the second case, the point $(0, \theta, 1 + \theta, 2 - \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is 0. In order to increase the objective function we must choose the first case, and we pick $\theta = 1$. We get the feasible solution $u_1 = (1, 0, 0, 3)$ corresponding to the basis (A^1, A^4), so it is a basic feasible solution, and the value of the objective function is 1.

The vectors A^2 and A^3 are given in terms of the basis (A^1, A^4) by

$$\begin{aligned} A^2 &= -A^1 \\ A^3 &= A^1 + A^4. \end{aligned}$$

Repeating the process with $u_1 = (1, 0, 0, 3)$, we get

$$\begin{aligned} b &= A^1 + 3A^4 - \theta A^2 + \theta A^2 \\ &= A^1 + 3A^4 - \theta(-A^1) + \theta A^2 \\ &= (1 + \theta)A^1 + \theta A^2 + 3A^4, \end{aligned}$$

and

$$\begin{aligned} b &= A^1 + 3A^4 - \theta A^3 + \theta A^3 \\ &= A^1 + 3A^4 - \theta(A^1 + A^4) + \theta A^3 \\ &= (1 - \theta)A^1 + \theta A^3 + (3 - \theta)A^4. \end{aligned}$$

In the first case, the point $(1 + \theta, \theta, 0, 3)$ is a feasible solution for all $\theta \geq 0$ and the value of the objective function if $1 + \theta$, and in the second case, the point $(1 - \theta, 0, \theta, 3 - \theta)$ is a

feasible solution iff $0 \leq \theta \leq 1$ and the value of the objective function is $1 - \theta$. This time, we are in the situation where the points

$$(1 + \theta, \theta, 0, 3) = (1, 0, 0, 3) + \theta(1, 1, 0, 0), \quad \theta \geq 0$$

form an infinite ray in the set of feasible solutions, and the objective function $1 + \theta$ is unbounded from above on this ray. This indicates that our linear program, although feasible, is unbounded.

Let us now describe a step of the simplex algorithm in general.

8.2 The Simplex Algorithm in General

We assume that we already have an initial vertex u_0 to start from. This vertex corresponds to a basic feasible solution with basis K_0. We will show later that it is always possible to find a basic feasible solution of a Linear Program (P) is standard form, or to detect that (P) has no feasible solution.

The idea behind the simplex algorithm is this: Given a pair (u, K) consisting of a basic feasible solution u and a basis K for u, find another pair (u^+, K^+) consisting of another basic feasible solution u^+ and a basis K^+ for u^+, such that K^+ is obtained from K by deleting some basic index $k^- \in K$ and adding some nonbasic index $j^+ \notin K$, in such a way that the value of the objective function increases (preferably strictly). The step which consists in swapping the vectors A^{k^-} and A^{j^+} is called a *pivoting step*.

Let u be a given vertex corresponds to a basic feasible solution with basis K. Since the m vectors A^k corresponding to indices $k \in K$ are linearly independent, they form a basis, so for every nonbasic $j \notin K$, we write

$$A^j = \sum_{k \in K} \gamma_k^j A^k. \tag{$*$}$$

We let $\gamma_K^j \in \mathbb{R}^m$ be the vector given by $\gamma_K^j = (\gamma_k^j)_{k \in K}$. Actually, since the vector γ_K^j depends on K, to be very precise we should denote its components by $(\gamma_K^j)_k$, but to simplify notation we usually write γ_k^j instead of $(\gamma_K^j)_k$ (unless confusion arises). We will explain later how the coefficients γ_k^j can be computed efficiently.

Since u is a feasible solution we have $u \geq 0$ and $Au = b$, that is,

$$\sum_{k \in K} u_k A^k = b. \tag{$**$}$$

For every nonbasic $j \notin K$, a candidate for entering the basis K, we try to find a new vertex $u(\theta)$ that improves the objective function, and for this we add $-\theta A^j + \theta A^j = 0$ to b in

Equation $(**)$ and then replace the occurrence of A^j in $-\theta A^j$ by the right hand side of Equation $(*)$ to obtain

$$\begin{aligned} b &= \sum_{k\in K} u_k A^k - \theta A^j + \theta A^j \\ &= \sum_{k\in K} u_k A^k - \theta\Big(\sum_{k\in K} \gamma_k^j A^k\Big) + \theta A^j \\ &= \sum_{k\in K} \Big(u_k - \theta\gamma_k^j\Big) A^k + \theta A^j. \end{aligned}$$

Consequently, the vector $u(\theta)$ appearing on the right-hand side of the above equation given by

$$u(\theta)_i = \begin{cases} u_i - \theta\gamma_i^j & \text{if } i \in K \\ \theta & \text{if } i = j \\ 0 & \text{if } i \notin K \cup \{j\} \end{cases}$$

automatically satisfies the constraints $Au(\theta) = b$, and this vector is a feasible solution iff

$$\theta \geq 0 \quad \text{and} \quad u_k \geq \theta\gamma_k^j \quad \text{for all } k \in K.$$

Obviously $\theta = 0$ is a solution, and if

$$\theta^j = \min\left\{ \frac{u_k}{\gamma_k^j} \;\middle|\; \gamma_k^j > 0,\ k \in K \right\} > 0,$$

then we have a range of feasible solutions for $0 \leq \theta \leq \theta^j$. The value of the objective function for $u(\theta)$ is

$$cu(\theta) = \sum_{k\in K} c_k(u_k - \theta\gamma_k^j) + \theta c_j = cu + \theta\Big(c_j - \sum_{k\in K} \gamma_k^j c_k\Big).$$

Since the potential change in the objective function is

$$\theta\Big(c_j - \sum_{k\in K} \gamma_k^j c_k\Big)$$

and $\theta \geq 0$, if $c_j - \sum_{k\in K} \gamma_k^j c_k \leq 0$, then the objective function can't be increased.

However, if $c_{j^+} - \sum_{k\in K} \gamma_k^{j^+} c_k > 0$ for some $j^+ \notin K$, and if $\theta^{j^+} > 0$, then the objective function can be strictly increased by choosing any $\theta > 0$ such that $\theta \leq \theta^{j^+}$, so it is natural to zero at least one coefficient of $u(\theta)$ by picking $\theta = \theta^{j^+}$, which also maximizes the increase of the objective function. In this case (Case below (B2)), we obtain a new feasible solution $u^+ = u(\theta^{j^+})$.

Now, if $\theta^{j^+} > 0$, then there is some index $k \in K$ such $u_k > 0$, $\gamma_k^{j^+} > 0$, and $\theta^{j^+} = u_k/\gamma_k^{j^+}$, so we can pick such an index k^- for the vector A^{k^-} leaving the basis K. We claim that

$K^+ = (K - \{k^-\}) \cup \{j^+\}$ is a basis. This is because the coefficient $\gamma_{k^-}^{j^+}$ associated with the column A^{k^-} is nonzero (in fact, $\gamma_{k^-}^{j^+} > 0$), so Equation $(*)$, namely

$$A^{j^+} = \gamma_{k^-}^{j^+} A^{k^-} + \sum_{k \in K - \{k^-\}} \gamma_k^{j^+} A^k,$$

yields the equation

$$A^{k^-} = (\gamma_{k^-}^{j^+})^{-1} A^{j^+} - \sum_{k \in K - \{k^-\}} (\gamma_{k^-}^{j^+})^{-1} \gamma_k^{j^+} A^k,$$

and these equations imply that the subspaces spanned by the vectors $(A^k)_{k \in K}$ and the vectors $(A^k)_{k \in K^+}$ are identical. However, K is a basis of dimension m so this subspace has dimension m, and since K^+ also has m elements, it must be a basis. Therefore, $u^+ = u(\theta^{j^+})$ is a basic feasible solution.

The above case is the most common one, but other situations may arise. In what follows, we discuss all eventualities.

Case (A).

We have $c_j - \sum_{k \in K} \gamma_k^j c_k \leq 0$ for all $j \notin K$. Then it turns out that u is an *optimal solution*. Otherwise, we are in Case (B).

Case (B).

We have $c_j - \sum_{k \in K} \gamma_k^j c_k > 0$ for some $j \notin K$ (not necessarily unique). There are three subcases.

Case (B1).

If for some $j \notin K$ as above we also have $\gamma_k^j \leq 0$ for all $k \in K$, since $u_k \geq 0$ for all $k \in K$, this places no restriction on θ, and the objective function is *unbounded above*. This is demonstrated by Example 8.3 with $K = \{3, 4\}$ and $j = 2$ since $\gamma_{\{3,4\}}^2 = (-1, 0)$.

Case (B2).

There is some index $j^+ \notin K$ such that simultaneously

(1) $c_{j^+} - \sum_{k \in K} \gamma_k^{j^+} c_k > 0$, which means that the objective function can potentially be increased;

(2) There is some $k \in K$ such that $\gamma_k^{j^+} > 0$, and for every $k \in K$, if $\gamma_k^{j^+} > 0$ then $u_k > 0$, which implies that $\theta^{j^+} > 0$.

If we pick $\theta = \theta^{j^+}$ where

$$\theta^{j^+} = \min\left\{ \frac{u_k}{\gamma_k^{j^+}} \;\middle|\; \gamma_k^{j^+} > 0,\ k \in K \right\} > 0,$$

then the feasible solution u^+ given by

$$u_i^+ = \begin{cases} u_i - \theta^{j^+}\gamma_i^{j^+} & \text{if } i \in K \\ \theta^{j^+} & \text{if } i = j^+ \\ 0 & \text{if } i \notin K \cup \{j^+\} \end{cases}$$

is a vertex of $\mathcal{P}(A, b)$. If we pick any index $k^- \in K$ such that $\theta^{j^+} = u_{k^-}/\gamma_{k^-}^{j+}$, then $K^+ = (K - \{k^-\}) \cup \{j^+\}$ is a basis for u^+. The vector A^{j^+} enters the new basis K^+, and the vector A^{k^-} leaves the old basis K. This is a *pivoting step*. The objective function increases strictly. This is demonstrated by Example 8.2 with $K = \{3, 4, 5\}$, $j = 1$, and $k = 4$, Then $\gamma^1_{\{3,4,5\}} = (-1, 1, 0)$, with $\gamma_4^1 = 1$. Since $u = (0, 0, 1, 3, 2)$, $\theta^1 = \frac{u_4}{\gamma_4^1} = 3$, and the new optimal solutions becomes $u^+ = (3, 0, 1 - 3(-1), 3 - 3(1), 2 - 3(0)) = (3, 0, 4, 0, 2)$.

Case (B3).

There is some index $j \notin K$ such that $c_j - \sum_{k\in K} \gamma_k^j c_k > 0$, and for each of the indices $j \notin K$ satisfying the above property we have simultaneously

(1) $c_j - \sum_{k\in K} \gamma_k^j c_k > 0$, which means that the objective function can potentially be increased;

(2) There is some $k \in K$ such that $\gamma_k^j > 0$, and $u_k = 0$, which implies that $\theta^j = 0$.

Consequently, the objective function *does not change*. In this case, u is a degenerate basic feasible solution.

We can associate to $u^+ = u$ a new basis K^+ as follows: Pick any index $j^+ \notin K$ such that

$$c_{j^+} - \sum_{k\in K} \gamma_k^{j^+} c_k > 0,$$

and any index $k^- \in K$ such that

$$\gamma_{k^-}^{j^+} > 0,$$

and let $K^+ = (K - \{k^-\}) \cup \{j^+\}$. As in Case (B2), The vector A^{j^+} enters the new basis K^+, and the vector A^{k^-} leaves the old basis K. This is a *pivoting step*. However, the objective function *does not change* since $\theta^{j+} = 0$. This is demonstrated by Example 8.1 with $K = \{3, 4\}$, $j = 2$, and $k = 3$.

It is easy to prove that in Case (A) the basic feasible solution u is an optimal solution, and that in Case (B1) the linear program is unbounded. We already proved that in Case (B2) the vector u^+ and its basis K^+ constitutes a basic feasible solution, and the proof in Case (B3) is similar. For details, see Ciarlet [19] (Chapter 10).

It is convenient to reinterpret the various cases considered by introducing the following sets:

$$B_1 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0,\ \max_{k \in K} \gamma_k^j \leq 0 \Big\}$$

$$B_2 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0,\ \max_{k \in K} \gamma_k^j > 0,\ \min\Big\{ \frac{u_k}{\gamma_k^j} \;\Big|\; k \in K, \gamma_k^j > 0 \Big\} > 0 \Big\}$$

$$B_3 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0,\ \max_{k \in K} \gamma_k^j > 0,\ \min\Big\{ \frac{u_k}{\gamma_k^j} \;\Big|\; k \in K, \gamma_k^j > 0 \Big\} = 0 \Big\},$$

and

$$B = B_1 \cup B_2 \cup B_3 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0 \Big\}.$$

Then it is easy to see that the following equivalences hold:

$$\begin{aligned} &\text{Case (A)} \iff B = \emptyset, \qquad \text{Case (B)} \iff B \neq \emptyset \\ &\text{Case (B1)} \iff B_1 \neq \emptyset \\ &\text{Case (B2)} \iff B_2 \neq \emptyset \\ &\text{Case (B3)} \iff B_3 \neq \emptyset. \end{aligned}$$

Furthermore, Cases (A) and (B), Cases (B1) and (B3), and Cases (B2) and (B3) are mutually exclusive, while Cases (B1) and (B2) are not.

If Case (B1) and Case (B2) arise simultaneously, we opt for Case (B1) which says that the Linear Program (P) is unbounded and terminate the algorithm.

Here are a few remarks about the method.

In Case (B2), which is the path followed by the algorithm most frequently, various choices have to be made for the index $j^+ \notin K$ for which $\theta^{j^+} > 0$ (the new index in K^+). Similarly, various choices have to be made for the index $k^- \in K$ leaving K, but such choices are typically less important.

Similarly in Case (B3), various choices have to be made for the new index $j^+ \notin K$ going into K^+. In Cases (B2) and (B3), criteria for making such choices are called *pivot rules*.

Case (B3) only arises when u is a degenerate vertex. But even if u is degenerate, Case (B2) may arise if $u_k > 0$ whenever $\gamma_k^j > 0$. It may also happen that u is nondegenerate but as a result of Case (B2), the new vertex u^+ is degenerate because at least two components $u_{k_1} - \theta^{j^+}\gamma_{k_1}^{j^+}$ and $u_{k_2} - \theta^{j^+}\gamma_{k_2}^{j^+}$ vanish for some distinct $k_1, k_2 \in K$.

Cases (A) and (B1) correspond to situations where the algorithm terminates, and Case (B2) can only arise a finite number of times during execution of the simplex algorithm, since the objective function is strictly increased from vertex to vertex and there are only finitely many vertices. Therefore, if the simplex algorithm is started on any initial basic feasible solution u_0, then one of three mutually exclusive situations may arise:

(1) There is a finite sequence of occurrences of Case (B2) and/or Case (B3) ending with an occurrence of Case (A). Then the last vertex produced by the algorithm is an optimal solution. This is what occurred in Examples 8.1 and 8.2.

(2) There is a finite sequence of occurrences of Case (B2) and/or Case (B3) ending with an occurrence of Case (B1). We conclude that the problem is unbounded, and thus has no solution. This is what occurred in Example 8.3.

(3) There is a finite sequence of occurrences of Case (B2) and/or Case (B3), followed by an infinite sequence of Case (B3). If this occurs, the algorithm visits the some basis twice. This a phenomenon known as *cycling*. In this eventually the algorithm fails to come to a conclusion.

There are examples for which cycling occur, although this is rare in practice. Such an example is given in Chvatal [18]; see Chapter 3, pages 31-32, for an example with seven variables and three equations that cycles after six iterations under a certain pivot rule.

The third possibility can be avoided by the choice of a suitable pivot rule. Two of these rules are *Bland's rule* and the *lexicographic rule*; see Chvatal [18] (Chapter 3, pages 34-38).

Bland's rule says: choose the smallest of the eligible incoming indices $j^+ \notin K$, and similarly choose the smallest of the eligible outgoing indices $k^- \in K$.

It can be proven that cycling cannot occur if Bland's rule is chosen as the pivot rule. The proof is very technical; see Chvatal [18] (Chapter 3, pages 37-38), Matousek and Gardner [42] (Chapter 5, Theorem 5.8.1), and Papadimitriou and Steiglitz [47] (Section 2.7). Therefore, assuming that some initial basic feasible solution is provided, and using a suitable pivot rule (such as Bland's rule), the simplex algorithm always terminates and either yields an optimal solution or reports that the linear program is unbounded. Unfortunately, Bland's rules is one of the slowest pivot rules.

The choice of a pivot rule affects greatly the number of pivoting steps that the simplex algorithms goes through. It is not our intention here to explain the various pivot rules. We simply mention the following rules, referring the reader to Matousek and Gardner [42] (Chapter 5, Section 5.7) or to the texts cited in Section 6.1.

1. Largest coefficient, or Dantzig's rule.

2. Largest increase.

3. Steepest edge.

4. Bland's Rule.

5. Random edge.

The steepest edge rule is one of the most popular. The idea is to maximize the ratio

$$\frac{c(u^+ - u)}{\|u^+ - u\|}.$$

The random edge rule picks the index $j^+ \notin K$ of the entering basis vector uniformly at random among all eligible indices.

Let us now return to the issue of the initialization of the simplex algorithm. We use the Linear Program $(\widehat{P})$ introduced during the proof of Theorem 7.7.

Consider a Linear Program $(P2)$

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax = b \text{ and } x \geq 0, \end{aligned}$$

in standard form where A is an $m \times n$ matrix of rank m.

First, observe that since the constraints are equations, we can ensure that $b \geq 0$, because every equation $a_i x = b_i$ where $b_i < 0$ can be replaced by $-a_i x = -b_i$. The next step is to introduce the Linear Program $(\widehat{P})$ in standard form

$$\begin{aligned} &\text{maximize} && -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} && \widehat{A}\,\widehat{x} = b \text{ and } \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ and $\widehat{x}$ are given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n+m} \end{pmatrix}.$$

Since we assumed that $b \geq 0$, the vector $\widehat{x} = (0_n, b)$ is a feasible solution of $(\widehat{P})$, in fact a basic feasible solutions since the matrix associated with the indices $n+1, \ldots, n+m$ is the identity matrix I_m. Furthermore, since $x_i \geq 0$ for all i, the objective function $-(x_{n+1} + \cdots + x_{n+m})$ is bounded above by 0.

If we execute the simplex algorithm with a pivot rule that prevents cycling, starting with the basic feasible solution $(0_n, d)$, since the objective function is bounded by 0, the simplex algorithm terminates with an optimal solution given by some basic feasible solution, say (u^*, w^*), with $u^* \in \mathbb{R}^n$ and $w^* \in \mathbb{R}^m$.

As in the proof of Theorem 7.7, for every feasible solution $u \in \mathcal{P}(A, b)$, the vector $(u, 0_m)$ is an optimal solution of $(\widehat{P})$. Therefore, if $w^* \neq 0$, then $\mathcal{P}(A, b) = \emptyset$, since otherwise for every feasible solution $u \in \mathcal{P}(A, b)$ the vector $(u, 0_m)$ would yield a value of the objective function $-(x_{n+1} + \cdots + x_{n+m})$ equal to 0, but (u^*, w^*) yields a strictly negative value since $w^* \neq 0$.

Otherwise, $w^* = 0$, and u^* is a feasible solution of $(P2)$. Since $(u^*, 0_m)$ is a basic feasible solution of $(\widehat{P})$ the columns corresponding to nonzero components of u^* are linearly independent. Some of the coordinates of u^* could be equal to 0, but since A has rank m we can add columns of A to obtain a basis K^* associated with u^*, and u^* is indeed a basic feasible solution of $(P2)$.

Running the simplex algorithm on the Linear Program $\widehat{P}$ to obtain an initial feasible solution (u_0, K_0) of the linear program $(P2)$ is called *Phase I* of the simplex algorithm. Running the simplex algorithm on the Linear Program $(P2)$ with some initial feasible solution (u_0, K_0) is called *Phase II* of the simplex algorithm. If a feasible solution of the Linear Program $(P2)$ is readily available then Phase I is skipped. Sometimes, at the end of Phase I, an optimal solution of $(P2)$ is already obtained.

In summary, we proved the following fact worth recording.

Proposition 8.1. *For any Linear Program* $(P2)$

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax = b \text{ and } x \geq 0, \end{aligned}$$

in standard form, where A is an $m \times n$ matrix of rank m and $b \geq 0$, consider the Linear Program $(\widehat{P})$ in standard form

$$\begin{aligned} &\text{maximize} && -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} && \widehat{A}\,\widehat{x} = b \text{ and } \widehat{x} \geq 0. \end{aligned}$$

The simplex algorithm with a pivot rule that prevents cycling started on the basic feasible solution $\widehat{x} = (0_n, b)$ of $(\widehat{P})$ terminates with an optimal solution (u^*, w^*).

(1) If $w^* \neq 0$, then $\mathcal{P}(A, b) = \emptyset$, that is, the Linear Program $(P2)$ has no feasible solution.

(2) If $w^* = 0$, then $\mathcal{P}(A, b) \neq \emptyset$, and u^* is a basic feasible solution of $(P2)$ associated with some basis K.

Proposition 8.1 shows that determining whether the polyhedron $\mathcal{P}(A, b)$ defined by a system of equations $Ax = b$ and inequalities $x \geq 0$ is nonempty is decidable. This decision procedure uses a fail-safe version of the simplex algorithm (that prevents cycling), and the proof that it always terminates and returns an answer is nontrivial.

8.3 How to Perform a Pivoting Step Efficiently

We now discuss briefly how to perform the computation of (u^+, K^+) from a basic feasible solution (u, K).

In order to avoid applying permutation matrices it is preferable to allow a basis K to be a sequence of indices, possibly out of order. Thus, for any $m \times n$ matrix A (with $m \leq n$) and any sequence $K = (k_1, k_2, \cdots, k_m)$ of m elements with $k_i \in \{1, \ldots, n\}$, the matrix A_K denotes the $m \times m$ matrix whose ith column is the k_ith column of A, and similarly for any vector $u \in \mathbb{R}^n$ (resp. any linear form $c \in (\mathbb{R}^n)^*$), the vector $u_K \in \mathbb{R}^m$ (the linear form $c_K \in (\mathbb{R}^m)^*$) is the vector whose ith entry is the k_ith entry in u (resp. the linear whose ith entry is the k_ith entry in c).

For each nonbasic $j \notin K$, we have

$$A^j = \gamma^j_{k_1} A^{k_1} + \cdots + \gamma^j_{k_m} A^{k_m} = A_K \gamma^j_K,$$

so the vector γ^j_K is given by $\gamma^j_K = A_K^{-1} A^j$, that is, by solving the system

$$A_K \gamma^j_K = A^j. \tag{$*_\gamma$}$$

To be very precise, since the vector γ^j_K depends on K its components should be denoted by $(\gamma^j_K)_{k_i}$, but as we said before, to simplify notation we write $\gamma^j_{k_i}$ instead of $(\gamma^j_K)_{k_i}$.

In order to decide which case applies ((A), (B1), (B2), (B3)), we need to compute the numbers $c_j - \sum_{k \in K} \gamma^j_k c_k$ for all $j \notin K$. For this, observe that

$$c_j - \sum_{k \in K} \gamma^j_k c_k = c_j - c_K \gamma^j_K = c_j - c_K A_K^{-1} A^j.$$

If we write $\beta_K = c_K A_K^{-1}$, then

$$c_j - \sum_{k \in K} \gamma^j_k c_k = c_j - \beta_K A^j,$$

and we see that $\beta_K^\top \in \mathbb{R}^m$ is the solution of the system $\beta_K^\top = (A_K^{-1})^\top c_k^\top$, which means that $\beta_K^\top$ is the solution of the system

$$A_K^\top \beta_K^\top = c_K^\top. \tag{$*_\beta$}$$

Remark: Observe that since u is a basis feasible solution of (P), we have $u_j = 0$ for all $j \notin K$, so u is the solution of the equation $A_K u_K = b$. As a consequence, the value of the objective function for u is $cu = c_K u_K = c_K A_K^{-1} b$. This fact will play a crucial role in Section 9.2 to show that when the simplex algorithm terminates with an optimal solution of the Linear Program (P), then it also produces an optimal solution of the Dual Linear Program (D).

Assume that we have a basic feasible solution u, a basis K for u, and that we also have the matrix A_K as well its inverse A_K^{-1} (perhaps implicitly) and also the inverse $(A_K^\top)^{-1}$ of $A_K^\top$ (perhaps implicitly). Here is a description of an iteration step of the simplex algorithm, following almost exactly Chvatal (Chvatal [18], Chapter 7, Box 7.1).

An Iteration Step of the (Revised) Simplex Method

Step 1. Compute the numbers $c_j - \sum_{k\in K} \gamma_k^j c_k = c_j - \beta_K A^j$ for all $j \notin K$, and for this, compute $\beta_K^\top$ as the solution of the system

$$A_K^\top \beta_K^\top = c_K^\top.$$

If $c_j - \beta_K A^j \leq 0$ for all $j \notin K$, stop and return the optimal solution u (Case (A)).

Step 2. If Case (B) arises, use a pivot rule to determine which index $j^+ \notin K$ should enter the new basis K^+ (the condition $c_{j^+} - \beta_K A^{j^+} > 0$ should hold).

Step 3. Compute $\max_{k\in K} \gamma_k^{j^+}$. For this, solve the linear system

$$A_K \gamma_K^{j^+} = A^{j^+}.$$

Step 4. If $\max_{k\in K} \gamma_k^{j^+} \leq 0$, then stop and report that Linear Program (P) is unbounded (Case (B1)).

Step 5. If $\max_{k\in K} \gamma_k^{j^+} > 0$, use the ratios $u_k/\gamma_k^{j^+}$ for all $k \in K$ such that $\gamma_k^{j^+} > 0$ to compute θ^{j^+}, and use a pivot rule to determine which index $k^- \in K$ such that $\theta^{j^+} = u_{k^-}/\gamma_{k^-}^{j+}$ should leave K (Case (B2)).

If $\max_{k\in K} \gamma_k^{j^+} = 0$, then use a pivot rule to determine which index k^- for which $\gamma_{k^-}^{j^+} > 0$ should leave the basis K (Case (B3)).

Step 6. Update u, K, and A_K, to u^+ and K^+, and A_{K^+}. During this step, given the basis K specified by the sequence $K = (k_1, \ldots, k_\ell, \ldots, k_m)$, with $k^- = k_\ell$, then K^+ is the sequence obtained by replacing k_ℓ by the incoming index j^+, so $K^+ = (k_1, \ldots, j^+, \ldots, k_m)$ with j^+ in the ℓth slot.

The vector u is easily updated. To compute A_{K^+} from A_K we take advantage of the fact that A_K and A_{K^+} only differ by a *single column*, namely the ℓth column A^{j^+}, which is given by the linear combination

$$A^{j^+} = A_K \gamma_K^{j^+}.$$

To simplify notation, denote $\gamma_K^{j^+}$ by γ, and recall that $k^- = k_\ell$. If $K = (k_1, \ldots, k_m)$, then $A_K = [A^{k_1} \cdots A^{k^-} \cdots A^{i_m}]$, and since A_{K^+} is the result of replacing the ℓth column A^{k^-} of A_K by the column A^{j^+}, we have

$$A_{K^+} = [A^{k_1} \cdots A^{j^+} \cdots A^{i_m}] = [A^{k_1} \cdots A_K\gamma \cdots A^{i_m}] = A_K E(\gamma),$$

where $E(\gamma)$ is the following invertible matrix obtained from the identity matrix I_m by re-

placing its ℓth column by γ:

$$E(\gamma) = \begin{pmatrix} 1 & & \gamma_1 & & \\ & \ddots & \vdots & & \\ & 1 & \gamma_{\ell-1} & & \\ & & \gamma_\ell & & \\ & & \gamma_{\ell+1} & 1 & \\ & & \vdots & & \ddots \\ & & \gamma_m & & 1 \end{pmatrix}.$$

Since $\gamma_\ell = \gamma_{k^-}^{j^+} > 0$, the matrix $E(\gamma)$ is invertible, and it is easy to check that its inverse is given by

$$E(\gamma)^{-1} = \begin{pmatrix} 1 & & -\gamma_\ell^{-1}\gamma_1 & & \\ & \ddots & \vdots & & \\ & 1 & -\gamma_\ell^{-1}\gamma_{\ell-1} & & \\ & & \gamma_\ell^{-1} & & \\ & & -\gamma_\ell^{-1}\gamma_{\ell+1} & 1 & \\ & & \vdots & & \ddots \\ & & -\gamma_\ell^{-1}\gamma_m & & 1 \end{pmatrix},$$

which is very cheap to compute. We also have

$$A_{K^+}^{-1} = E(\gamma)^{-1} A_K^{-1}.$$

Consequently, if A_K and A_K^{-1} are available, then A_{K^+} and $A_{K^+}^{-1}$ can be computed cheaply in terms of A_K and A_K^{-1} and matrices of the form $E(\gamma)$. Then the systems $(*_\gamma)$ to find the vectors γ_K^j can be solved cheaply.

Since

$$A_{K^+}^\top = E(\gamma)^\top A_K^\top$$

and

$$(A_{K^+}^\top)^{-1} = (A_K^\top)^{-1}(E(\gamma)^\top)^{-1},$$

the matrices $A_{K^+}^\top$ and $(A_{K^+}^\top)^{-1}$ can also be computed cheaply from $A_K^\top$, $(A_K^\top)^{-1}$, and matrices of the form $E(\gamma)^\top$. Thus the systems $(*_\beta)$ to find the linear forms β_K can also be solved cheaply.

A matrix of the form $E(\gamma)$ is called an *eta matrix*; see Chvatal [18] (Chapter 7). We showed that the matrix A_{K^s} obtained after s steps of the simplex algorithm can be written as

$$A_{K^s} = A_{K^{s-1}} E_s$$

for some eta matrix E_s, so A_{k^s} can be written as the product

$$A_{K^s} = E_1 E_2 \cdots E_s$$

of s eta matrices. Such a factorization is called an *eta factorization*. The eta factorization can be used to either invert A_{K^s} or to solve a system of the form $A_{K_s}\gamma = A^{j^+}$ iteratively. Which method is more efficient depends on the sparsity of the E_i.

In summary, there are cheap methods for finding the next basic feasible solution (u^+, K^+) from (u, K). We simply wanted to give the reader a flavor of these techniques. We refer the reader to texts on linear programming for detailed presentations of methods for implementing efficiently the simplex method. In particular, the *revised simplex method* is presented in Chvatal [18], Papadimitriou and Steiglitz [47], Bertsimas and Tsitsiklis [10], and Vanderbei [66].

8.4 The Simplex Algorithm Using Tableaux

We now describe a formalism for presenting the simplex algorithm, namely *(full) tableaux*. This is the traditional formalism used in all books, modulo minor variations. A particularly nice feature of the tableau formalism is that the update of a tableau can be performed using elementary row operations *identical* to the operations used during the reduction of a matrix to row reduced echelon form (rref). What differs is the criterion for the choice of the pivot.

Since the quantities $c_j - c_K\gamma_K^j$ play a crucial role in determining which column A^j should come into the basis, the notation $\overline{c}_j$ is used to denote $c_j - c_K\gamma_K^j$, which is called the *reduced cost* of the variable x_j. The reduced costs actually depend on K so to be very precise we should denote them by $(\overline{c}_K)_j$, but to simplify notation we write $\overline{c}_j$ instead of $(\overline{c}_K)_j$. We will see shortly how $(\overline{c}_{K^+})_i$ is computed in terms of $(\overline{c}_K)_i$.

Observe that the data needed to execute the next step of the simplex algorithm are

(1) The current basic solution u_K and its basis $K = (k_1, \ldots, k_m)$.

(2) The reduced costs $\overline{c}_j = c_j - c_K A_K^{-1} A^j = c_j - c_K\gamma_K^j$, for all $j \notin K$.

(3) The vectors $\gamma_K^j = (\gamma_{k_i}^j)_{i=1}^m$ for all $j \notin K$, that allow us to express each A^j as $A_K\gamma_K^j$.

All this information can be packed into a $(m+1) \times (n+1)$ matrix called a *(full) tableau* organized as follows:

$c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ_1^1	$\cdots$	γ_1^j	$\cdots$	γ_1^n
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ_m^1	$\cdots$	γ_m^j	$\cdots$	γ_m^n

It is convenient to think as the first row as Row 0, and of the first column as Column 0. Row 0 contains the current value of the objective function and the reduced costs. Column 0, except for its top entry, contains the components of the current basic solution u_K, and

the remaining columns, except for their top entry, contain the vectors γ_K^j. Observe that the γ_K^j corresponding to indices j in K constitute a permutation of the identity matrix I_m. The entry $\gamma_{k^-}^{j^+}$ is called the *pivot* element. A tableau together with the new basis $K^+ = (K - \{k^-\}) \cup \{j^+\}$ contains all the data needed to compute the new u_{K^+}, the new $\gamma_{K^+}^j$, and the new reduced costs $(\overline{c}_{K^+})_j$.

If we define the $m \times n$ matrix Γ as the matrix $\Gamma = [\gamma_K^1 \ \cdots \ \gamma_K^n]$ whose jth column is γ_K^j, and $\overline{c}$ as the row vector $\overline{c} = (\overline{c}_1 \ \cdots \ \overline{c}_n)$, then the above tableau is denoted concisely by

$c_K u_K$	$\overline{c}$
u_K	Γ

We now show that the update of a tableau can be performed using elementary row operations identical to the operations used during the reduction of a matrix to row reduced echelon form (rref).

If $K = (k_1, \ldots, k_m)$, j^+ is the index of the incoming basis vector, $k^- = k_\ell$ is the index of the column leaving the basis, and if $K^+ = (k_1, \ldots, k_{\ell-1}, j^+, k_{\ell+1}, \ldots, k_m)$, since $A_{K^+} = A_K E(\gamma_K^{j^+})$, the new columns $\gamma_{K^+}^j$ are computed in terms of the old columns γ_K^j using $(*_\gamma)$ and the equations

$$\gamma_{K^+}^j = A_{K^+}^{-1} A^j = E(\gamma_K^{j^+})^{-1} A_K^{-1} A^j = E(\gamma_K^{j^+})^{-1} \gamma_K^j.$$

Consequently, the matrix Γ^+ is given in terms of Γ by

$$\Gamma^+ = E(\gamma_K^{j^+})^{-1} \Gamma.$$

But the matrix $E(\gamma_K^{j^+})^{-1}$ is of the form

$$E(\gamma_K^{j^+})^{-1} = \begin{pmatrix} 1 & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_1}^{j^+} & & \\ & \ddots & \vdots & & \\ & & 1 \quad -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_{\ell-1}}^{j^+} & & \\ & & (\gamma_{k^-}^{j^+})^{-1} & & \\ & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_{\ell+1}}^{j^+} \quad 1 & & \\ & & \vdots & \ddots & \\ & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_m}^{j^+} & & 1 \end{pmatrix},$$

with the column involving the γs in the ℓth column, and Γ^+ is obtained by applying the following elementary row operations to Γ:

1. Multiply Row ℓ by $1/\gamma_{k^-}^{j^+}$ (the inverse of the pivot) to make the entry on Row ℓ and Column j^+ equal to 1.

2. Subtract $\gamma_{k_i}^{j^+} \times$ (the normalized) Row ℓ from Row i, for $i = 1, \ldots, \ell - 1, \ell + 1, \ldots, m$.

These are *exactly* the elementary row operations that reduce the ℓth column $\gamma_K^{j^+}$ of Γ to the ℓth column of the identity matrix I_m. Thus, this step is identical to the sequence of steps that the procedure to convert a matrix to row reduced echelon from executes on the ℓth column of the matrix. The only difference is the criterion for the choice of the pivot.

Since the new basic solution u_{K^+} is given by $u_{K^+} = A_{K^+}^{-1} b$, we have

$$u_{K^+} = E(\gamma_K^{j^+})^{-1} A_K^{-1} b = E(\gamma_K^{j^+})^{-1} u_K.$$

This means that u_+ is obtained from u_K by applying exactly the *same* elementary row operations that were applied to Γ. Consequently, just as in the procedure for reducing a matrix to rref, we can apply elementary row operations to the matrix $[u_k \ \Gamma]$, which consists of rows $1, \ldots, m$ of the tableau.

Once the new matrix Γ^+ is obtained, the new reduced costs are given by the following proposition.

Proposition 8.2. *Given any Linear Program $(P2)$ in standard form*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix of rank m, if (u, K) is a basic (not necessarily feasible) solution of $(P2)$ and if $K^+ = (K - \{k^-\}) \cup \{j^+\}$, with $K = (k_1, \ldots, k_m)$ and $k^- = k_\ell$, then for $i = 1, \ldots, n$ we have

$$c_i - c_{K^+}\gamma_{K^+}^i = c_i - c_K \gamma_K^i - \frac{\gamma_{k^-}^i}{\gamma_{k^-}^{j^+}}(c_{j^+} - c_K \gamma_K^{j^+}).$$

Using the reduced cost notation, the above equation is

$$(\overline{c}_{K^+})_i = (\overline{c}_K)_i - \frac{\gamma_{k^-}^i}{\gamma_{k^-}^{j^+}}(\overline{c}_K)_{j^+}.$$

Proof. Without any loss of generality and to simplify notation assume that $K = (1, \ldots, m)$ and write j for j^+ and ℓ for k_m. Since $\gamma_K^i = A_K^{-1} A^i$, $\gamma_{K^+}^i = A_{K^+}^{-1} A^i$, and $A_{K^+} = A_K E(\gamma_K^j)$, we have

$$c_i - c_{K^+}\gamma_{K^+}^i = c_i - c_{K^+} A_{K^+}^{-1} A^i = c_i - c_{K^+} E(\gamma_K^j)^{-1} A_K^{-1} A^i = c_i - c_{K^+} E(\gamma_K^j)^{-1} \gamma_K^i,$$

where

$$E(\gamma_K^j)^{-1} = \begin{pmatrix} 1 & & & -(\gamma_\ell^j)^{-1}\gamma_1^j & & \\ & \ddots & & \vdots & & \\ & & 1 & -(\gamma_\ell^j)^{-1}\gamma_{\ell-1}^j & & \\ & & & (\gamma_\ell^j)^{-1} & & \\ & & & -(\gamma_\ell^j)^{-1}\gamma_{\ell+1}^j & 1 & \\ & & & \vdots & & \ddots \\ & & & -(\gamma_\ell^j)^{-1}\gamma_m^j & & 1 \end{pmatrix}$$

where the ℓth column contains the γs. Since $c_{K^+} = (c_1, \ldots, c_{\ell-1}, c_j, c_{\ell+1}, \ldots, c_m)$, we have

$$c_{K^+}E(\gamma_K^j)^{-1} = \left(c_1, \ldots, c_{\ell-1}, \frac{c_j}{\gamma_\ell^j} - \sum_{k=1,k\neq\ell}^{m} c_k\frac{\gamma_k^j}{\gamma_\ell^j}, c_{\ell+1}, \ldots, c_m\right),$$

and

$$\begin{aligned}
c_{K^+}E(\gamma_K^j)^{-1}\gamma_K^i &= \begin{pmatrix} c_1 & \ldots & c_{\ell-1} & \dfrac{c_j}{\gamma_\ell^j} - \displaystyle\sum_{k=1,k\neq\ell}^{m} c_k\frac{\gamma_k^j}{\gamma_\ell^j} & c_{\ell+1} & \ldots & c_m \end{pmatrix} \begin{pmatrix} \gamma_1^i \\ \vdots \\ \gamma_{\ell-1}^i \\ \gamma_\ell^i \\ \gamma_{\ell+1}^i \\ \vdots \\ \gamma_m^i \end{pmatrix} \\
&= \sum_{k=1,k\neq\ell}^{m} c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j - \sum_{k=1,k\neq\ell}^{m} c_k\gamma_k^j\right) \\
&= \sum_{k=1,k\neq\ell}^{m} c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j + c_\ell\gamma_\ell^j - \sum_{k=1}^{m} c_k\gamma_k^j\right) \\
&= \sum_{k=1}^{m} c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j - \sum_{k=1}^{m} c_k\gamma_k^j\right) \\
&= c_K\gamma_K^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}(c_j - c_K\gamma_K^j),
\end{aligned}$$

and thus

$$c_i - c_{K^+}\gamma_{K^+}^i = c_i - c_{K^+}E(\gamma_K^j)^{-1}\gamma_K^i = c_i - c_K\gamma_K^i - \frac{\gamma_\ell^i}{\gamma_\ell^j}(c_j - c_K\gamma_K^j),$$

as claimed. □

Since $(\gamma_{k^-}^1, \ldots, \gamma_{k^-}^n)$ is the ℓth row of Γ, we see that Proposition 8.2 shows that

$$\overline{c}_{K^+} = \overline{c}_K - \frac{(\overline{c}_K)_{j^+}}{\gamma_{k^-}^{j^+}}\,\Gamma_\ell, \tag{$\dagger$}$$

where Γ_ℓ denotes the ℓ-th row of Γ and $\gamma_{k^-}^{j^+}$ is the pivot. This means that $\overline{c}_{K^+}$ is obtained by the elementary row operations which consist of first normalizing the ℓth row by dividing it by the pivot $\gamma_{k^-}^{j^+}$, and then subtracting $(\overline{c}_K)_{j^+}\times$ the normalized Row ℓ from $\overline{c}_K$. *These are exactly the row operations that make the reduced cost $(\overline{c}_K)_{j^+}$ zero.*

Remark: It easy easy to show that we also have

$$\overline{c}_{K^+} = c - c_{K^+}\Gamma^+.$$

We saw in Section 8.2 that the change in the objective function after a pivoting step during which column j^+ comes in and column k^- leaves is given by

$$\theta^{j^+}\left(c_{j^+} - \sum_{k\in K} \gamma_k^{j^+} c_k\right) = \theta^{j^+}(\overline{c}_K)_{j^+},$$

where

$$\theta^{j^+} = \frac{u_{k^-}}{\gamma_{k^-}^{j^+}}.$$

If we denote the value of the objective function $c_K u_K$ by z_K, then we see that

$$z_{K^+} = z_K + \frac{(\overline{c}_K)_{j^+}}{\gamma_{k^-}^{j^+}}\, u_{k^-}.$$

This means that the new value z_{K^+} of the objective function is obtained by first normalizing the ℓth row by dividing it by the pivot $\gamma_{k^-}^{j^+}$, and then adding $(\overline{c}_K)_{j^+}\times$ the zeroth entry of the normalized ℓth line by $(\overline{c}_K)_{j^+}$ to the zeroth entry of line 0.

In updating the reduced costs, we subtract rather than add $(\overline{c}_K)_{j^+}\times$ the normalized row ℓ from row 0. This suggests storing $-z_K$ as the zeroth entry on line 0 rather than z_K, because then all the entries row 0 are updated by the *same* elementary row operations. Therefore, from now on, we use tableau of the form

$-c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ_1^1	$\cdots$	γ_1^j	$\cdots$	γ_1^n
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ_m^1	$\cdots$	γ_m^j	$\cdots$	γ_m^n

The simplex algorithm first chooses the incoming column j^+ by picking some column for which $\overline{c}_j > 0$, and then chooses the outgoing column k^- by considering the ratios $u_k/\gamma_k^{j^+}$ for which $\gamma_k^{j^+} > 0$ (along column j^+), and picking k^- to achieve the minimum of these ratios.

Here is an illustration of the simplex algorithm using elementary row operations on an example from Papadimitriou and Steiglitz [47] (Section 2.9).

Example 8.4. Consider the linear program

$$\begin{aligned}
&\text{maximize} && -2x_2 - x_4 - 5x_7 \\
&\text{subject to} && \\
& && x_1 + x_2 + x_3 + x_4 = 4 \\
& && x_1 + x_5 = 2 \\
& && x_3 + x_6 = 3 \\
& && 3x_2 + x_3 + x_7 = 6 \\
& && x_1, x_2, x_3, x_4, x_5, x_6, x_7 \geq 0.
\end{aligned}$$

We have the basic feasible solution $u = (0, 0, 0, 4, 2, 3, 6)$, with $K = (4, 5, 6, 7)$. Since $c_K = (-1, 0, 0, -5)$ and $c = (0, -2, 0, -1, 0, 0 - 5)$ the first tableau is

34	1	14	6	0	0	0	0
$u_4 = 4$	1	1	1	1	0	0	0
$u_5 = 2$	(1)	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

Since $\bar{c}_j = c_j - c_K \gamma_K^j$, Row 0 is obtained by subtracting $-1\times$ Row 1 and $-5\times$ Row 4 from $c = (0, -2, 0, -1, 0, 0, -5)$. Let us pick Column $j^+ = 1$ as the incoming column. We have the ratios (for positive entries on Column 1)

$$4/1,\ 2/1,$$

and since the minimum is 2, we pick the outgoing column to be Column $k^- = 5$. The pivot 1 is indicated in red. The new basis is $K = (4, 1, 6, 7)$. Next we apply row operations to reduce Column 1 to the second vector of the identity matrix I_4. For this, we subtract Row 2 from Row 1. We get the tableau

34	1	14	6	0	0	0	0
$u_4 = 2$	0	1	1	1	-1	0	0
$u_1 = 2$	(1)	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

To compute the new reduced costs, we want to set $\bar{c}_1$ to 0, so we apply the identical row operations and subtract Row 2 from Row 0 to obtain the tableau

32	0	14	6	0	-1	0	0
$u_4 = 2$	0	1	(1)	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

Next, pick Column $j^+ = 3$ as the incoming column. We have the ratios (for positive entries on Column 3)

$$2/1,\ 3/1,\ 6/1,$$

and since the minimum is 2, we pick the outgoing column to be Column $k^- = 4$. The pivot 1 is indicated in red and the new basis is $K = (3, 1, 6, 7)$. Next we apply row operations to reduce Column 3 to the first vector of the identity matrix I_4. For this, we subtract Row 1 from Row 3 and from Row 4 and obtain the tableau:

32	0	14	6	0	-1	0	0
$u_3 = 2$	0	1	(1)	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 1$	0	-1	0	-1	1	1	0
$u_7 = 4$	0	2	0	-1	1	0	1

To compute the new reduced costs, we want to set $\overline{c}_3$ to 0, so we subtract $6\times$ Row 1 from Row 0 to get the tableau

20	0	8	0	-6	5	0	0
$u_3 = 2$	0	(1)	1	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 1$	0	-1	0	-1	1	1	0
$u_7 = 4$	0	2	0	-1	1	0	1

Next we pick $j^+ = 2$ as the incoming column. We have the ratios (for positive entries on Column 2)

$$2/1,\ 4/2,$$

and since the minimum is 2, we pick the outgoing column to be Column $k^- = 3$. The pivot 1 is indicated in red and the new basis is $K = (2, 1, 6, 7)$. Next we apply row operations to reduce Column 2 to the first vector of the identity matrix I_4. For this, we add Row 1 to Row 3 and subtract $2\times$ Row 1 from Row 4 to obtain the tableau:

20	0	8	0	-6	5	0	0
$u_2 = 2$	0	(1)	1	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 0$	0	0	-2	-3	3	0	1

To compute the new reduced costs, we want to set $\overline{c}_2$ to 0, so we subtract $8\times$ Row 1 from Row 0 to get the tableau

4	0	0	−8	−14	13	0	0
$u_2 = 2$	0	1	1	1	−1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 0$	0	0	−2	−3	3	0	1

The only possible incoming column corresponds to $j^+ = 5$. We have the ratios (for positive entries on Column 5)

$$2/1,\ 0/3,$$

and since the minimum is 0, we pick the outgoing column to be Column $k^- = 7$. The pivot 3 is indicated in red and the new basis is $K = (2, 1, 6, 5)$. Since the minimum is 0, the basis $K = (2, 1, 6, 5)$ is degenerate (indeed, the component corresponding to the index 5 is 0). Next we apply row operations to reduce Column 5 to the fourth vector of the identity matrix I_4. For this, we multiply Row 4 by 1/3, and then add the normalized Row 4 to Row 1 and subtract the normalized Row 4 from Row 2 to obtain the tableau:

4	0	0	−8	−14	13	0	0
$u_2 = 2$	0	1	1/3	0	0	0	1/3
$u_1 = 2$	1	0	2/3	1	0	0	−1/3
$u_6 = 3$	0	0	1	0	0	1	0
$u_5 = 0$	0	0	−2/3	−1	1	0	1/3

To compute the new reduced costs, we want to set $\overline{c}_5$ to 0, so we subtract 13× Row 4 from Row 0 to get the tableau

4	0	0	2/3	−1	0	0	−13/3
$u_2 = 2$	0	1	1/3	0	0	0	1/3
$u_1 = 2$	1	0	2/3	1	0	0	−1/3
$u_6 = 3$	0	0	1	0	0	1	0
$u_5 = 0$	0	0	−2/3	−1	1	0	1/3

The only possible incoming column corresponds to $j^+ = 3$. We have the ratios (for positive entries on Column 3)

$$2/(1/3) = 6,\ 2/(2/3) = 3,\ 3/1 = 3,$$

and since the minimum is 3, we pick the outgoing column to be Column $k^- = 1$. The pivot 2/3 is indicated in red and the new basis is $K = (2, 3, 6, 5)$. Next we apply row operations to reduce Column 3 to the second vector of the identity matrix I_4. For this, we multiply Row 2 by 3/2, subtract (1/3)× (normalized Row 2) from Row 1, and subtract normalized Row 2 from Row 3, and add (2/3)× (normalized Row 2) to Row 4 to obtain the tableau:

4	0	0	2/3	−1	0	0	−13/3
$u_2 = 1$	−1/2	1	0	−1/2	0	0	1/2
$u_3 = 3$	3/2	0	(1)	3/2	0	0	−1/2
$u_6 = 0$	−3/2	0	0	−3/2	0	1	1/2
$u_5 = 2$	1	0	0	0	1	0	0

To compute the new reduced costs, we want to set $\overline{c}_3$ to 0, so we subtract (2/3)× Row 2 from Row 0 to get the tableau

2	−1	0	0	−2	0	0	−4
$u_2 = 1$	−1/2	1	0	−1/2	0	0	1/2
$u_3 = 3$	3/2	0	1	3/2	0	0	−1/2
$u_6 = 0$	−3/2	0	0	−3/2	0	1	1/2
$u_5 = 2$	1	0	0	0	1	0	0

Since all the reduced cost are ≤ 0, we have reached an optimal solution, namely $(0, 1, 3, 0, 2, 0, 0, 0)$, with optimal value -2.

The progression of the simplex algorithm from one basic feasible solution to another corresponds to the visit of vertices of the polyhedron $\mathcal{P}$ associated with the constraints of the linear program illustrated in Figure 8.4.

As a final comment, if it is necessary to run Phase I of the simplex algorithm, in the event that the simplex algorithm terminates with an optimal solution $(u^*, 0_m)$ and a basis K^* such that some $u_i = 0$, then the basis K^* contains indices of basic columns A^j corresponding to slack variables that need to be *driven out* of the basis. This is easy to achieve by performing a pivoting step involving some other column j^+ corresponding to one of the original variables (not a slack variable) for which $(\gamma_{K^*})_i^{j^+} \neq 0$. In such a step, it doesn't matter whether $(\gamma_{K^*})_i^{j^+} < 0$ or $(\overline{c}_{K^*})_{j^+} \leq 0$. If the original matrix A has no redundant equations, such a step is always possible. Otherwise, $(\gamma_{K^*})_i^j = 0$ for all non-slack variables, so we detected that the ith equation is redundant and we can delete it.

Other presentations of the tableau method can be found in Bertsimas and Tsitsiklis [10] and Papadimitriou and Steiglitz [47].

8.5 Computational Efficiency of the Simplex Method

Let us conclude with a few comments about the efficiency of the simplex algorithm. In *practice*, it was observed by Dantzig that for linear programs with $m < 50$ and $m + n < 200$, the simplex algorithms typically requires less than $3m/2$ iterations, but at most $3m$ iterations. This fact agrees with more recent empirical experiments with much larger programs that show that the number iterations is bounded by $3m$. Thus, it was somewhat of a shock in

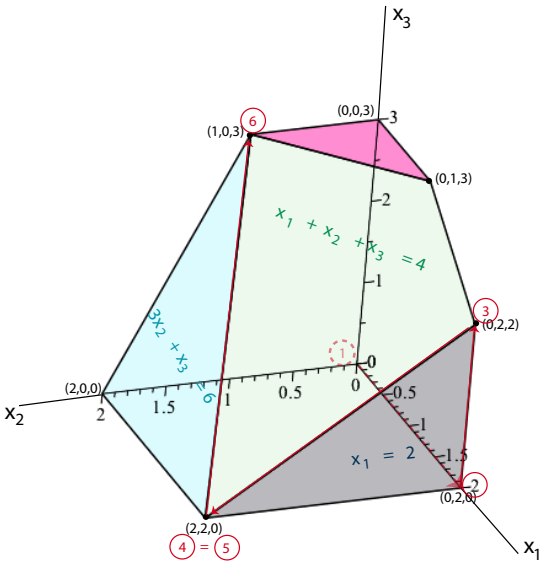


Figure 8.4: The polytope $\mathcal{P}$ associated with the linear program optimized by the tableau method. The red arrowed path traces the progression of the simplex method from the origin to the vertex $(0, 1, 3)$.

1972 when Klee and Minty found a linear program with n variables and n equations for which the simplex algorithm with Dantzig's pivot rule requires requires $2^n - 1$ iterations. This program (taken from Chvatal [18], page 47) is reproduced below:

$$\begin{aligned} &\text{maximize} \quad \sum_{j=1}^{n} 10^{n-j} x_j \\ &\text{subject to} \\ &\left(2 \sum_{j=1}^{i-1} 10^{i-j} x_j\right) + x_i \leq 100^{i-1} \\ &x_j \geq 0, \end{aligned}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n$.

If $p = \max(m, n)$, then, in terms of worse case behavior, for all currently known pivot rules, the simplex algorithm has exponential complexity in p. However, as we said earlier, in practice, nasty examples such as the Klee–Minty example seem to be rare, and the number of iterations appears to be linear in m.

Whether or not a pivot rule (a clairvoyant rule) for which the simplex algorithms runs in polynomial time in terms of m is still an *open problem*.

The *Hirsch conjecture* claims that there is some pivot rule such that the simplex algorithm finds an optimal solution in $O(p)$ steps. The best bound known so far due to Kalai and Kleitman is $m^{1+\ln n} = (2n)^{\ln m}$. For more on this topic, see Matousek and Gardner [42] (Section 5.9) and Bertsimas and Tsitsiklis [10] (Section 3.7).

Researchers have investigated the problem of finding upper bounds on the expected number of pivoting steps if a randomized pivot rule is used. Bounds better than 2^m (but of course, not polynomial) have been found.

Understanding the complexity of linear programing, in particular of the simplex algorithm, is still ongoing. The interested reader is referred to Matousek and Gardner [42] (Chapter 5, Section 5.9) for some pointers.

In the next section we consider important theoretical criteria for determining whether a set of constraints $Ax \leq b$ and $x \geq 0$ has a solution or not.

8.6 Summary

The main concepts and results of this chapter are listed below:

- Degenerate and nondegenerate basic feasible solution.
- Pivoting step.
- Pivot rule.
- Cycling.
- Bland's rule, Dantzig's rule, steepest edge rule, random edge rule, largest increase rule, lexicographic rule.
- Phase I and Phase II of the simplex algorithm.
- eta matrix, eta factorization.
- Revised simplex method.
- Reduced cost.
- Full tableaux.
- The Hirsch conjecture.

8.7 Problems

Problem 8.1. In Section 8.2 prove that if Case (A) arises, then the basic feasible solution u is an optimal solution. Prove that if Case (B1) arises, then the linear program is unbounded. Prove that if Case (B3) arises, then (u^+, K^+) is a basic feasible solution.

Problem 8.2. In Section 8.2 prove that the following equivalences hold:

$$\begin{aligned}
&\text{Case (A)} \Longleftrightarrow B = \emptyset, \qquad \text{Case (B)} \Longleftrightarrow B \neq \emptyset\\
&\text{Case (B1)} \Longleftrightarrow B_1 \neq \emptyset\\
&\text{Case (B2)} \Longleftrightarrow B_2 \neq \emptyset\\
&\text{Case (B3)} \Longleftrightarrow B_3 \neq \emptyset.
\end{aligned}$$

Furthermore, prove that Cases (A) and (B), Cases (B1) and (B3), and Cases (B2) and (B3) are mutually exclusive, while Cases (B1) and (B2) are not.

Problem 8.3. Consider the linear program (due to E.M.L. Beale):

$$\begin{aligned}
\text{maximize} \quad & (3/4)x_1 - 150x_2 + (1/50)x_3 - 6x_4\\
\text{subject to} \quad & \\
& (1/4)x_1 - 60x_2 - (1/25)x_3 + 9x_4 \leq 0\\
& (1/4)x_1 - 90x_2 - (1/50)x_3 + 3x_4 \leq 0\\
& \qquad\qquad\qquad\qquad\qquad\quad x_3 \leq 1\\
& x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0.
\end{aligned}$$

(1) Convert the above program to standard form.

(2) Show that if we apply the simplex algorithm with the pivot rule which selects the column entering the basis as the column of smallest index, then the method cycles.

Problem 8.4. Read carefully the proof given by Chvatal that the lexicographic pivot rule and Bland's pivot rule prevent cycling; see Chvatal [18] (Chapter 3, pages 34-38).

Problem 8.5. Solve the following linear program (from Chvatal [18], Chapter 3, page 44) using the two-phase simplex algorithm:

$$\begin{aligned}
\text{maximize} \quad & 3x_1 + x_2\\
\text{subject to} \quad & \\
& x_1 - x_2 \leq -1\\
& -x_1 - x_2 \leq -3\\
& 2x_1 + x_2 \leq 4\\
& x_1 \geq 0,\ x_2 \geq 0.
\end{aligned}$$

Problem 8.6. Solve the following linear program (from Chvatal [18], Chapter 3, page 44) using the two-phase simplex algorithm:

$$\begin{aligned} &\text{maximize} && 3x_1 + x_2 \\ &\text{subject to} && \\ & && x_1 - x_2 \leq -1 \\ & && -x_1 - x_2 \leq -3 \\ & && 2x_1 + x_2 \leq 2 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Problem 8.7. Solve the following linear program (from Chvatal [18], Chapter 3, page 44) using the two-phase simplex algorithm:

$$\begin{aligned} &\text{maximize} && 3x_1 + x_2 \\ &\text{subject to} && \\ & && x_1 - x_2 \leq -1 \\ & && -x_1 - x_2 \leq -3 \\ & && 2x_1 - x_2 \leq 2 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Problem 8.8. Show that the following linear program (from Chvatal [18], Chapter 3, page 43) is unbounded.

$$\begin{aligned} &\text{maximize} && x_1 + 3x_2 - x_3 \\ &\text{subject to} && \\ & && 2x_1 + 2x_2 - x_3 \leq 10 \\ & && 3x_1 - 2x_2 + x_3 \leq 10 \\ & && x_1 - 3x_2 + x_3 \leq 10 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0. \end{aligned}$$

Hint. Try $x_1 = 0, x_3 = t$, and a suitable value for x_2.

Chapter 9

Linear Programming and Duality

9.1 Variants of the Farkas Lemma

This section overlaps Section 4.1, but we believe that most readers will not mind if we review the versions of Farkas lemma that will be used to prove the strong duality theorem of linear programming. To avoid a clash with the versions of Farkas I and Farkas II from Section 4.1, we label the versions of Farkas lemma as Farkas Ib and Farkas IIb, even though Farkas IIb is the same as Farkas IIIb(3) (see Proposition 4.18).

If A is an $m \times n$ matrix and if $b \in \mathbb{R}^m$ is a vector, it is known from linear algebra that the linear system $Ax = b$ has no solution iff there is some linear form $y \in (\mathbb{R}^m)^*$ such that $yA = 0$ and $yb \neq 0$. This means that the linear from y vanishes on the columns $A^1, \ldots, A^n$ of A but does not vanish on b. Since the linear form y defines the linear hyperplane H of equation $yz = 0$ (with $z \in \mathbb{R}^m$), geometrically the equation $Ax = b$ has no solution iff there is a linear hyperplane H containing $A^1, \ldots, A^n$ and not containing b. This is a kind of separation theorem that says that the vectors $A^1, \ldots, A^n$ and b can be separated by some linear hyperplane H.

What we would like to do is to generalize this kind of criterion, first to a system $Ax = b$ subject to the constraints $x \geq 0$, and next to sets of inequality constraints $Ax \leq b$ and $x \geq 0$. There are indeed such criteria going under the name of *Farkas lemma*.

The key is a separation result involving polyhedral cones known as the Farkas–Minkowski proposition. We have the following fundamental separation lemma, which is just a restatement of Proposition 4.15.

Proposition 9.1. *Let $C \subseteq \mathbb{R}^n$ be a closed nonempty cone. For any point $a \in \mathbb{R}^n$, if $a \notin C$, then there is a linear hyperplane H (through 0) such that*

1. *C lies in one of the two half-spaces determined by H.*
2. *$a \notin H$*
3. *a lies in the other half-space determined by H.*

We say that H strictly separates C and a.

The Farkas–Minkowski proposition is Proposition 9.1 applied to a polyhedral cone

$$C = \{\lambda_1 a_1 + \cdots + \lambda_n a_n \mid \lambda_i \geq 0,\ i = 1, \ldots, n\}$$

where $\{a_1, \ldots, a_n\}$ is a *finite* number of vectors $a_i \in \mathbb{R}^n$. By Proposition 4.13, any polyhedral cone is closed, so Proposition 9.1 applies and we obtain the following separation lemma.

Proposition 9.2. *(Farkas–Minkowski) Let $C \subseteq \mathbb{R}^n$ be a nonempty polyhedral cone $C = \text{cone}(\{a_1, \ldots, a_n\})$. For any point $b \in \mathbb{R}^n$, if $b \notin C$, then there is a linear hyperplane H (through 0) such that*

1. *C lies in one of the two half-spaces determined by H.*
2. *$a \notin H$*
3. *a lies in the other half-space determined by H.*

Equivalently, there is a nonzero linear form $y \in (\mathbb{R}^n)^*$ such that

1. *$ya_i \geq 0$ for $i = 1, \ldots, n$.*
2. *$yb < 0$.*

A direct proof of the Farkas–Minkowski proposition not involving Proposition 9.1 is given at the end of this section.

Remark: There is a generalization of the Farkas–Minkowski proposition applying to infinite dimensional real Hilbert spaces; see Ciarlet [19], Chapter 9.

Proposition 9.2 implies our first version of Farkas' lemma.

Proposition 9.3. *(Farkas Lemma, Version Ib) Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. The linear system $Ax = b$ has no solution $x \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $yA \geq 0_n^\top$ and $yb < 0$.*

Proof. First, assume that there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $yA \geq 0$ and $yb < 0$. If $x \geq 0$ is a solution of $Ax = b$, then we get

$$yAx = yb,$$

but if $yA \geq 0$ and $x \geq 0$, then $yAx \geq 0$, and yet by hypothesis $yb < 0$, a contradiction.

Next assume that $Ax = b$ has no solution $x \geq 0$. This means that b does not belong to the polyhedral cone $C = \text{cone}(\{A^1, \ldots, A^n\})$ spanned by the columns of A. By Proposition 9.2, there is a nonzero linear form $y \in (\mathbb{R}^m)^*$ such that

1. $yA^j \geq 0$ for $j = 1, \ldots, n$.

2. $yb < 0$,

which says that $yA \geq 0_n^\top$ and $yb < 0$. □

Next consider the solvability of a system of inequalities of the form $Ax \leq b$ and $x \geq 0$.

Proposition 9.4. *(Farkas Lemma, Version IIb) Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. The system of inequalities $Ax \leq b$ has no solution $x \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $y \geq 0_m^\top$, $yA \geq 0_n^\top$, and $yb < 0$.*

Proof. We use the trick of linear programming which consists of adding "slack variables" z_i to convert inequalities $a_i x \leq b_i$ into equations $a_i x + z_i = b_i$ with $z_i \geq 0$. If we let $z = (z_1, \ldots, z_m)$, it is obvious that the system $Ax \leq b$ has a solution $x \geq 0$ iff the equation

$$\begin{pmatrix} A & I_m \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix} = b$$

has a solution $\begin{pmatrix} x \\ z \end{pmatrix}$ with $x \geq 0$ and $z \geq 0$. Now by Farkas Ib, the above system has no solution with with $x \geq 0$ and $z \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that

$$y \begin{pmatrix} A & I_m \end{pmatrix} \geq 0_{n+m}^\top$$

and $yb < 0$, that is, $yA \geq 0_n^\top$, $y \geq 0_m^\top$, and $yb < 0$. □

In the next section we use Farkas IIb to prove the duality theorem in linear programming. Observe that by taking the negation of the equivalence in Farkas IIb we obtain a criterion of solvability, namely:

The system of inequalities $Ax \leq b$ has a solution $x \geq 0$ iff for every nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $y \geq 0_m^\top$, if $yA \geq 0_n^\top$, then $yb \geq 0$.

We now prove the Farkas–Minkowski proposition without using Proposition 3.3. This approach uses a basic property of the distance function from a point to a closed set.

Let $X \subseteq \mathbb{R}^n$ be any nonempty set and let $a \in \mathbb{R}^n$ be any point. The *distance* $d(a, X)$ from a to X is defined as

$$d(a, X) = \inf_{x \in X} \|a - x\|.$$

Here, $\| \ \|$ denotes the Euclidean norm.

Proposition 9.5. *Let $X \subseteq \mathbb{R}^n$ be any nonempty set and let $a \in \mathbb{R}^n$ be any point. If X is closed, then there is some $z \in X$ such that $\|a - z\| = d(a, X)$.*

Proof. Since X is nonempty, pick any $x_0 \in X$, and let $r = \|a - x_0\|$. If $B_r(a)$ is the closed ball $B_r(a) = \{x \in \mathbb{R}^n \mid \|x - a\| \leq r\}$, then clearly

$$d(a, X) = \inf_{x \in X} \|a - x\| = \inf_{x \in X \cap B_r(a)} \|a - x\| .$$

Since $B_r(a)$ is compact and X is closed, $K = X \cap B_r(a)$ is also compact. But the function $x \mapsto \|a - x\|$ defined on the compact set K is continuous, and the image of a compact set by a continuous function is compact, so by Heine–Borel it has a minimum that is achieved by some $z \in K \subseteq X$. □

Remark: If U is a nonempty, closed and convex subset of a Hilbert space V, a standard result of Hilbert space theory (the projection theorem) asserts that for any $v \in V$ there is a *unique* $p \in U$ such that

$$\|v - p\| = \inf_{u \in U} \|v - u\| = d(v, U),$$

and

$$\langle p - v, u - p \rangle \geq 0 \quad \text{for all } u \in U.$$

Here $\|w\| = \sqrt{\langle w, w \rangle}$, where $\langle -, - \rangle$ is the inner product of the Hilbert space V.

We can now give a proof of the Farkas–Minkowski proposition (Proposition 9.2).

Proof of the Farkas–Minkowski proposition. Let $C = \text{cone}(\{a_1, \ldots, a_m\})$ be a polyhedral cone (nonempty) and assume that $b \notin C$. By Proposition 4.13, the polyhedral cone is closed, and by Proposition 9.5 there is some $z \in C$ such that $d(b, C) = \|b - z\|$; that is, z is a point of C closest to b. Since $b \notin C$ and $z \in C$ we have $u = z - b \neq 0$, and we claim that the linear hyperplane H orthogonal to u does the job, as illustrated in Figure 9.1.

First let us show that

$$\langle u, z \rangle = \langle z - b, z \rangle = 0. \tag{$*_1$}$$

This is trivial if $z = 0$, so assume $z \neq 0$. If $\langle u, z \rangle \neq 0$, then either $\langle u, z \rangle > 0$ or $\langle u, z \rangle < 0$. In either case we show that we can find some point $z' \in C$ closer to b than z is, a contradiction.

Case 1: $\langle u, z \rangle > 0$.

Let $z' = (1 - \alpha)z$ for any α such that $0 < \alpha < 1$. Then $z' \in C$ and since $u = z - b$

$$z' - b = (1 - \alpha)z - (z - u) = u - \alpha z,$$

so

$$\|z' - b\|^2 = \|u - \alpha z\|^2 = \|u\|^2 - 2\alpha\langle u, z \rangle + \alpha^2 \|z\|^2 .$$

If we pick $\alpha > 0$ such that $\alpha < 2\langle u, z \rangle / \|z\|^2$, then $-2\alpha\langle u, z \rangle + \alpha^2 \|z\|^2 < 0$, so $\|z' - b\|^2 < \|u\|^2 = \|z - b\|^2$, contradicting the fact that z is a point of C closest to b.

Case 2: $\langle u, z \rangle < 0$.

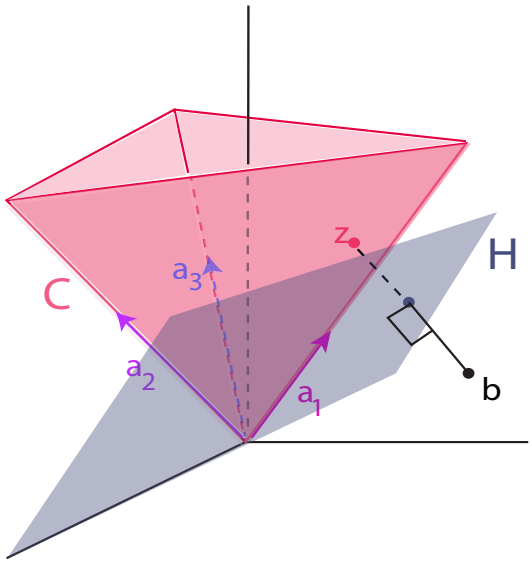


Figure 9.1: The hyperplane H, perpendicular to $z - b$, separates the point b from $C = \text{cone}(\{a_1, a_2, a_3\})$.

Let $z' = (1 + \alpha)z$ for any α such that $\alpha \geq -1$. Then $z' \in C$ and since $u = z - b$ we have $z' - b = (1 + \alpha)z - (z - u) = u + \alpha z$ so

$$\|z' - b\|^2 = \|u + \alpha z\|^2 = \|u\|^2 + 2\alpha\langle u, z\rangle + \alpha^2 \|z\|^2,$$

and if

$$0 < \alpha < -2\langle u, z\rangle / \|z\|^2,$$

then $2\alpha\langle u, z\rangle + \alpha^2 \|z\|^2 < 0$, so $\|z' - b\|^2 < \|u\|^2 = \|z - b\|^2$, a contradiction as above.

Therefore $\langle u, z\rangle = 0$. We have

$$\langle u, u\rangle = \langle u, z - b\rangle = \langle u, z\rangle - \langle u, b\rangle = -\langle u, b\rangle,$$

and since $u \neq 0$, we have $\langle u, u\rangle > 0$, so $\langle u, u\rangle = -\langle u, b\rangle$ implies that

$$\langle u, b\rangle < 0. \tag{$*_2$}$$

It remains to prove that $\langle u, a_i\rangle \geq 0$ for $i = 1, \ldots, m$. Pick any $x \in C$ such that $x \neq z$. We claim that

$$\langle b - z, x - z\rangle \leq 0. \tag{$*_3$}$$

Otherwise $\langle b - z, x - z\rangle > 0$, that is, $\langle z - b, x - z\rangle < 0$, and we show that we can find some point $z' \in C$ on the line segment $[z, x]$ closer to b than z is.

For any α such that $0 \leq \alpha \leq 1$, we have $z' = (1-\alpha)z + \alpha x = z + \alpha(x - z) \in C$, and since $z' - b = z - b + \alpha(x - z)$ we have

$$\|z' - b\|^2 = \|z - b + \alpha(x - z)\|^2 = \|z - b\|^2 + 2\alpha\langle z - b, x - z\rangle + \alpha^2 \|x - z\|^2,$$

so for any $\alpha > 0$ such that

$$\alpha < -2\langle z - b, x - z\rangle / \|x - z\|^2,$$

we have $2\alpha\langle z - b, x - z\rangle + \alpha^2 \|x - z\|^2 < 0$, which implies that $\|z' - b\|^2 < \|z - b\|^2$, contradicting that z is a point of C closest to b.

Since $\langle b - z, x - z\rangle \leq 0$, $u = z - b$, and by $(*_1)$ $\langle u, z\rangle = 0$, we have

$$0 \geq \langle b - z, x - z\rangle = \langle -u, x - z\rangle = -\langle u, x\rangle + \langle u, z\rangle = -\langle u, x\rangle,$$

which means that

$$\langle u, x\rangle \geq 0 \quad \text{for all } x \in C, \tag{$*_3$}$$

as claimed. In particular,

$$\langle u, a_i\rangle \geq 0 \quad \text{for } i = 1, \ldots, m. \tag{$*_4$}$$

Then, by $(*_2)$ and $(*_4)$, the linear form defined by $y = u^\top$ satisfies the properties $yb < 0$ and $ya_i \geq 0$ for $i = 1, \ldots, m$, which proves the Farkas–Minkowski proposition. □

There are other ways of proving the Farkas–Minkowski proposition, for instance using minimally infeasible systems or Fourier–Motzkin elimination; see Matousek and Gardner [42] (Chapter 6, Sections 6.6 and 6.7).

9.2 The Duality Theorem in Linear Programming

Let (P) be the linear program

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A a $m \times n$ matrix, and assume that (P) has a feasible solution and is bounded above. Since by hypothesis the objective function $x \mapsto cx$ is bounded on $\mathcal{P}(A, b)$, it might be useful to deduce an *upper bound* for cx from the inequalities $Ax \leq b$, for any $x \in \mathcal{P}(A, b)$. We can do this as follows: for every inequality

$$a_i x \leq b_i \quad 1 \leq i \leq m,$$

pick a nonnegative scalar y_i, multiply both sides of the above inequality by y_i obtaining

$$y_i a_i x \leq y_i b_i \quad 1 \leq i \leq m,$$

(the direction of the inequality is preserved since $y_i \geq 0$), and then add up these m equations, which yields

$$(y_1 a_1 + \cdots + y_m a_m)x \leq y_1 b_1 + \cdots + y_m b_m.$$

If we can pick the $y_i \geq 0$ such that

$$c \leq y_1 a_1 + \cdots + y_m a_m,$$

then since $x_j \geq 0$, we have

$$cx \leq (y_1 a_1 + \cdots + y_m a_m)x \leq y_1 b_1 + \cdots + y_m b_m,$$

namely we found an upper bound of the value cx of the objective function of (P) for any feasible solution $x \in \mathcal{P}(A, b)$. If we let y be the linear form $y = (y_1, \ldots, y_m)$, then since

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$$

$y_1 a_1 + \cdots + y_m a_m = yA$, and $y_1 b_1 + \cdots + y_m b_m = yb$, what we did was to look for some $y \in (\mathbb{R}^m)^*$ such that

$$c \leq yA, \quad y \geq 0,$$

so that we have

$$cx \leq yb \quad \text{for all } x \in \mathcal{P}(A, b). \tag{$*$}$$

Then it is natural to look for a "best" value of yb, namely a minimum value, which leads to the definition of the *dual* of the linear program (P), a notion due to John von Neumann.

Definition 9.1. Given any Linear Program (P)

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A an $m \times n$ matrix, the *dual* (D) of (P) is the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c \text{ and } y \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$.

The variables $y_1, \ldots, y_m$ are called the *dual variables*. The original Linear Program (P) is called the *primal* linear program and the original variables $x_1, \ldots, x_n$ are the *primal variables*.

Here is an explicit example of a linear program and its dual.

Example 9.1. Consider the linear program illustrated by Figure 9.2

$$\begin{aligned} &\text{maximize} && 2x_1 + 3x_2 \\ &\text{subject to} && \\ & && 4x_1 + 8x_2 \leq 12 \\ & && 2x_1 + x_2 \leq 3 \\ & && 3x_1 + 2x_2 \leq 4 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Its dual linear program is illustrated in Figure 9.3

$$\begin{aligned} &\text{minimize} && 12y_1 + 3y_2 + 4y_3 \\ &\text{subject to} && \\ & && 4y_1 + 2y_2 + 3y_3 \geq 2 \\ & && 8y_1 + y_2 + 2y_3 \geq 3 \\ & && y_1 \geq 0,\ y_2 \geq 0,\ y_3 \geq 0. \end{aligned}$$

It can be checked that $(x_1, x_2) = (1/2, 5/4)$ is an optimal solution of the primal linear program, with the maximum value of the objective function $2x_1 + 3x_2$ equal to $19/4$, and that $(y_1, y_2, y_3) = (5/16, 0, 1/4)$ is an optimal solution of the dual linear program, with the minimum value of the objective function $12y_1 + 3y_2 + 4y_3$ also equal to $19/4$.

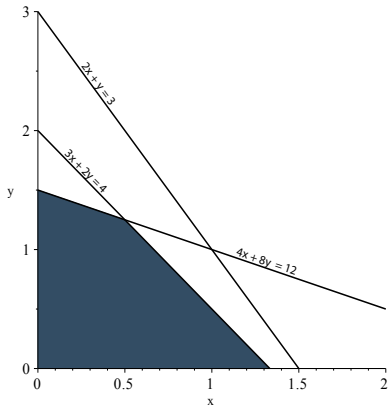


Figure 9.2: The $\mathcal{H}$-polytope for the linear program of Example 9.1. Note $x_1 \to x$ and $x_2 \to y$.

Observe that in the Primal Linear Program (P), we are looking for a *vector* $x \in \mathbb{R}^n$ maximizing the form cx, and that the constraints are determined by the action of the *rows* of the matrix A on x. On the other hand, in the Dual Linear Program (D), we are looking for a *linear form* $y \in (\mathbb{R}^*)^m$ minimizing the form yb, and the constraints are determined by

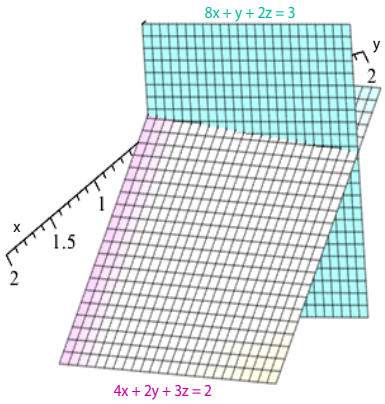


Figure 9.3: The $\mathcal{H}$-polyhedron for the dual linear program of Example 9.1 is the spacial region "above" the pink plane and in "front" of the blue plane. Note $y_1 \to x$, $y_2 \to y$, and $y_3 \to z$.

the action of y on the *columns* of A. This is the sense in which (D) is the *dual* (P). In most presentations, the fact that (P) and (D) perform a search for a solution in spaces that are dual to each other is obscured by excessive use of transposition.

To convert the Dual Program (D) to a standard maximization problem we change the objective function yb to $-b^\top y^\top$ and the inequality $yA \geq c$ to $-A^\top y^\top \leq -c^\top$. The Dual Linear Program (D) is now stated as (D')

$$\begin{aligned} &\text{maximize} && -b^\top y^\top \\ &\text{subject to} && -A^\top y^\top \leq -c^\top \text{ and } y^\top \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. Observe that the dual in maximization form (D'') of the Dual Program (D') gives back the Primal Program (P).

The above discussion established the following inequality known as *weak duality*.

Proposition 9.6. *(Weak Duality) Given any Linear Program* (P)

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A *an* $m \times n$ *matrix, for any feasible solution* $x \in \mathbb{R}^n$ *of the Primal Problem* (P) *and every feasible solution* $y \in (\mathbb{R}^m)^*$ *of the Dual Problem* (D)*, we have*

$$cx \leq yb.$$

Definition 9.2. We say that the Dual Linear Program (D) is *bounded below* if $\{yb \mid y^\top \in \mathcal{P}(-A^\top, -c^\top)\}$ is bounded below.

What happens if x^* is an optimal solution of (P) and if y^* is an optimal solution of (D)? We have $cx^* \leq y^*b$, but is there a "duality gap," that is, is it possible that $cx^* < y^*b$?

The answer is **no**, this is the *strong duality theorem*. Actually, the strong duality theorem asserts more than this.

Theorem 9.7. *(Strong Duality for Linear Programming) Let (P) be any linear program*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A an $m \times n$ matrix. The Primal Problem (P) has a feasible solution and is bounded above iff the Dual Problem (D) has a feasible solution and is bounded below. Furthermore, if (P) has a feasible solution and is bounded above, then for every optimal solution x^* of (P) and every optimal solution y^* of (D), we have

$$cx^* = y^*b.$$

Proof. If (P) has a feasible solution and is bounded above, then we know from Proposition 7.1 that (P) has some optimal solution. Let x^* be any optimal solution of (P). First we will show that (D) has a feasible solution v.

Let $\mu = cx^*$ be the maximum of the objective function $x \mapsto cx$. Then for any $\epsilon > 0$, the system of inequalities

$$Ax \leq b, \quad x \geq 0, \quad cx \geq \mu + \epsilon$$

has no solution, since otherwise μ would not be the maximum value of the objective function cx. We would like to apply Farkas II, so first we transform the above system of inequalities into the system

$$\begin{pmatrix} A \\ -c \end{pmatrix} x \leq \begin{pmatrix} b \\ -(\mu + \epsilon) \end{pmatrix}.$$

By Proposition 4.16 (Farkas II), there is some linear form $(\lambda, z) \in (\mathbb{R}^{m+1})^*$ such that $\lambda \geq 0$, $z \geq 0$,

$$\begin{pmatrix} \lambda & z \end{pmatrix} \begin{pmatrix} A \\ -c \end{pmatrix} \geq 0_m^\top,$$

and

$$\begin{pmatrix} \lambda & z \end{pmatrix} \begin{pmatrix} b \\ -(\mu + \epsilon) \end{pmatrix} < 0,$$

which means that

$$\lambda A - zc \geq 0_m^\top, \quad \lambda b - z(\mu + \epsilon) < 0,$$

that is,

$$\begin{aligned} \lambda A &\geq zc \\ \lambda b &< z(\mu + \epsilon) \\ \lambda &\geq 0, \ \ z \geq 0. \end{aligned}$$

On the other hand, since $x^* \geq 0$ is an optimal solution of the system $Ax \leq b$, by Farkas II again (by taking the negation of the equivalence), since $\lambda A \geq 0$ (for the same λ as before), we must have

$$\lambda b \geq 0. \tag{$*_1$}$$

We claim that $z > 0$. Otherwise, since $z \geq 0$, we must have $z = 0$, but then

$$\lambda b < z(\mu + \epsilon)$$

implies

$$\lambda b < 0, \tag{$*_2$}$$

and since $\lambda b \geq 0$ by $(*_1)$, we have a contradiction. Consequently, we can divide by $z > 0$ without changing the direction of inequalities, and we obtain

$$\begin{aligned} \frac{\lambda}{z} A &\geq c \\ \frac{\lambda}{z} b &< \mu + \epsilon \\ \frac{\lambda}{z} &\geq 0, \end{aligned}$$

which shows that $v = \lambda/z$ is a feasible solution of the Dual Problem (D). However, weak duality (Proposition 9.6) implies that $cx^* = \mu \leq yb$ for any feasible solution $y \geq 0$ of the Dual Program (D), so (D) is bounded below and by Proposition 7.1 applied to the version of (D) written as a maximization problem, we conclude that (D) has some optimal solution. For any optimal solution y^* of (D), since v is a feasible solution of (D) such that $vb < \mu + \epsilon$, we must have

$$\mu \leq y^* b < \mu + \epsilon,$$

and since our reasoning is valid for *any* $\epsilon > 0$, we conclude that $cx^* = \mu = y^* b$.

If we assume that the dual program (D) has a feasible solution and is bounded below, since the dual of (D) is (P), we conclude that (P) is also feasible and bounded above. □

The strong duality theorem can also be proven by the simplex method, because when it terminates with an optimal solution of (P), the final tableau also produces an optimal solution y of (D) that can be read off the reduced costs of columns $n + 1, \ldots, n + m$ by flipping their signs. We follow the proof in Ciarlet [19] (Chapter 10).

Theorem 9.8. *Consider the Linear Program (P),*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

its equivalent version $(P2)$ *in standard form,*

$$\begin{aligned} &\text{maximize} \quad \widehat{c}\,\widehat{x} \\ &\text{subject to} \quad \widehat{A}\widehat{x} = b \text{ and } \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ *is an* $m \times (n+m)$ *matrix,* $\widehat{c}$ *is a linear form in* $(\mathbb{R}^{n+m})^*$*, and* $\widehat{x} \in \mathbb{R}^{n+m}$*, given by*

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{c} = \begin{pmatrix} c & 0_m^\top \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \overline{x} = \begin{pmatrix} x_{n+1} \\ \vdots \\ x_{n+m} \end{pmatrix}, \quad \widehat{x} = \begin{pmatrix} x \\ \overline{x} \end{pmatrix},$$

and the Dual (D) *of* (P) *given by*

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c \text{ and } y \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$*. If the simplex algorithm applied to the Linear Program* $(P2)$ *terminates with an optimal solution* $(\widehat{u}^*, K^*)$*, where* $\widehat{u}^*$ *is a basic feasible solution and* K^* *is a basis for* $\widehat{u}^*$*, then* $y^* = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}$ *is an optimal solution for* (D) *such that* $\widehat{c}\,\widehat{u}^* = y^*b$*. Furthermore,* y^* *is given in terms of the reduced costs by* $y^* = -((\overline{c}_{K^*})_{n+1} \ \ldots (\overline{c}_{K^*})_{n+m})$*.*

Proof. We know that K^* is a subset of $\{1, \ldots, n+m\}$ consisting of m indices such that the corresponding columns of $\widehat{A}$ are linearly independent. Let $N^* = \{1, \ldots, n+m\} - K^*$. The simplex method terminates with an optimal solution in Case (A), namely when

$$\widehat{c}_j - \sum_{k \in k} \gamma_k^j \widehat{c}_k \leq 0 \quad \text{for all } j \in N^*,$$

where $\widehat{A}^j = \sum_{k \in K^*} \gamma_k^j \widehat{A}^k$, or using the notations of Section 8.3,

$$\widehat{c}_j - \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}^j \leq 0 \quad \text{for all } j \in N^*.$$

The above inequalities can be written as

$$\widehat{c}_{N^*} - \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}_{N^*} \leq 0_n^\top,$$

or equivalently as

$$\widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}_{N^*} \geq \widehat{c}_{N^*}. \tag{$*_1$}$$

The value of the objective function for the optimal solution $\widehat{u}^*$ is $\widehat{c}\,\widehat{u}^* = \widehat{c}_{K^*}\widehat{u}^*_{K^*}$, and since $\widehat{u}^*_{K^*}$ satisfies the equation $\widehat{A}_{K^*}\widehat{u}^*_{K^*} = b$, the value of the objective function is

$$\widehat{c}_{K^*}\,\widehat{u}^*_{K^*} = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}b. \tag{$*_2$}$$

Then if we let $y^* = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}$, obviously we have $y^*b = \widehat{c}_{K^*}\widehat{u}_{K^*}$, so if we can prove that y^* is a feasible solution of the Dual Linear program (D), by weak duality, y^* is an optimal solution of (D). We have

$$y^*\widehat{A}_{K^*} = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}_{K^*} = \widehat{c}_{K^*}, \tag{$*_3$}$$

and by $(*_1)$ we get

$$y^*\widehat{A}_{N^*} = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}_{N^*} \geq \widehat{c}_{N^*}. \tag{$*_4$}$$

Let P be the $(n+m)\times(n+m)$ permutation matrix defined so that

$$\widehat{A}\,P = \begin{pmatrix} A & I_m \end{pmatrix} P = \begin{pmatrix} \widehat{A}_{K^*} & \widehat{A}_{N^*} \end{pmatrix}.$$

Then we also have

$$\widehat{c}\,P = \begin{pmatrix} c & 0_m^\top \end{pmatrix} P = \begin{pmatrix} \widehat{c}_{K^*} & \widehat{c}_{N^*} \end{pmatrix}.$$

Using Equations $(*_3)$ and $(*_4)$ we obtain

$$y^*\begin{pmatrix} \widehat{A}_{K^*} & \widehat{A}_{N^*} \end{pmatrix} \geq \begin{pmatrix} \widehat{c}_{K^*} & \widehat{c}_{N^*} \end{pmatrix},$$

that is,

$$y^*\begin{pmatrix} A & I_m \end{pmatrix} P \geq \begin{pmatrix} c & 0_m^\top \end{pmatrix} P,$$

which is equivalent to

$$y^*\begin{pmatrix} A & I_m \end{pmatrix} \geq \begin{pmatrix} c & 0_m^\top \end{pmatrix},$$

that is

$$y^*A \geq c, \quad y \geq 0,$$

and these are exactly the conditions that say that y^* is a feasible solution of the Dual Program (D).

The reduced costs are given by $(\widehat{c}_{K^*})_i = \widehat{c}_i - \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}^i$, for $i = 1, \ldots, n+m$. But for $i = n+j$ with $j = 1, \ldots, m$ each column $\widehat{A}^{n+j}$ is the jth vector of the identity matrix I_m and by definition $\widehat{c}_{n+j} = 0$, so

$$(\widehat{c}_{K^*})_{n+j} = -(\widehat{c}_{K^*}\widehat{A}_{K^*}^{-1})_j = -y_j^* \quad j = 1, \ldots, m,$$

as claimed. □

The fact that the above proof is fairly short is deceptive because this proof relies on the fact that there are versions of the simplex algorithm using pivot rules that prevent cycling, but the proof that such pivot rules work correctly is quite lengthy. Other proofs are given in Matousek and Gardner [42] (Chapter 6, Sections 6.3), Chvatal [18] (Chapter 5), and Papadimitriou and Steiglitz [47] (Section 2.7).

Observe that since the last m rows of the final tableau are actually obtained by multipling $[u\ \widehat{A}]$ by $\widehat{A}_{K^*}^{-1}$, the $m \times m$ matrix consisting of the last m columns and last m rows of the final

tableau is $\widehat{A}_{K^*}^{-1}$ (basically, the simplex algorithm has performed the steps of a Gauss–Jordan reduction). This fact allows saving some steps in the primal dual method.

By combining weak duality and strong duality, we obtain the following theorem which shows that exactly four cases arise.

Theorem 9.9. *(Duality Theorem of Linear Programming) Let (P) be any linear program*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

and let (D) be its dual program

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c \text{ and } y \geq 0, \end{aligned}$$

with A an $m \times n$ matrix. Then exactly one of the following possibilities occur:

(1) Neither (P) nor (D) has a feasible solution.

(2) (P) is unbounded and (D) has no feasible solution.

(3) (P) has no feasible solution and (D) is unbounded.

(4) Both (P) and (D) have a feasible solution. Then both have an optimal solution, and for every optimal solution x^* of (P) and every optimal solution y^* of (D), we have

$$cx^* = y^*b.$$

An interesting corollary of Theorem 9.9 is that there is a test to determine whether a Linear Program (P) has an optimal solution.

Corollary 9.10. *The Primal Program (P) has an optimal solution iff the following set of constraints is satisfiable:*

$$\begin{aligned} Ax &\leq b \\ yA &\geq c \\ cx &\geq yb \\ x \geq 0,\ y &\geq 0_m^\top. \end{aligned}$$

In fact, for any feasible solution (x^*, y^*) of the above system, x^* is an optimal solution of (P) and y^* is an optimal solution of (D)

9.3 Complementary Slackness Conditions

Another useful corollary of the strong duality theorem is the following result known as the *equilibrium theorem*.

Theorem 9.11. *(Equilibrium Theorem) For any Linear Program (P) and its Dual Linear Program (D) (with set of inequalities $Ax \leq b$ where A is an $m \times n$ matrix, and objective function $x \mapsto cx$), for any feasible solution x of (P) and any feasible solution y of (D), x and y are optimal solutions iff*

$$y_i = 0 \qquad \text{for all } i \text{ for which } \sum_{j=1}^{n} a_{ij}x_j < b_i \tag{$*_D$}$$

and

$$x_j = 0 \qquad \text{for all } j \text{ for which } \sum_{i=1}^{m} y_i a_{ij} > c_j. \tag{$*_P$}$$

Proof. First assume that $(*_D)$ and $(*_P)$ hold. The equations in $(*_D)$ say that $y_i = 0$ unless $\sum_{j=1}^{n} a_{ij}x_j = b_i$, hence

$$yb = \sum_{i=1}^{m} y_i b_i = \sum_{i=1}^{m} y_i \sum_{j=1}^{n} a_{ij}x_j = \sum_{i=1}^{m}\sum_{j=1}^{n} y_i a_{ij} x_j.$$

Similarly, the equations in $(*_P)$ say that $x_j = 0$ unless $\sum_{i=1}^{m} y_i a_{ij} = c_j$, hence

$$cx = \sum_{j=1}^{n} c_j x_j = \sum_{j=1}^{n}\sum_{i=1}^{m} y_i a_{ij} x_j.$$

Consequently, we obtain

$$cx = yb.$$

By weak duality (Proposition 9.6), we have

$$cx \leq yb = cx$$

for all feasible solutions x of (P), so x is an optimal solution of (P). Similarly,

$$yb = cx \leq yb$$

for all feasible solutions y of (D), so y is an optimal solution of (D).

Let us now assume that x is an optimal solution of (P) and that y is an optimal solution of (D). Then as in the proof of Proposition 9.6,

$$\sum_{j=1}^{n} c_j x_j \leq \sum_{i=1}^{m}\sum_{j=1}^{n} y_i a_{ij} x_j \leq \sum_{i=1}^{m} y_i b_i.$$

By strong duality, since x and y are optimal solutions the above inequalities are actually equalities, so in particular we have

$$\sum_{j=1}^{n}\Big(c_j - \sum_{i=1}^{m} y_i a_{ij}\Big)x_j = 0.$$

Since x and y are feasible, $x_i \geq 0$ and $y_j \geq 0$, so if $\sum_{i=1}^{m} y_i a_{ij} > c_j$, we must have $x_j = 0$. Similarly, we have

$$\sum_{i=1}^{m} y_i\Big(\sum_{j=1}^{m} a_{ij}x_j - b_i\Big) = 0,$$

so if $\sum_{j=1}^{m} a_{ij}x_j < b_i$, then $y_i = 0$. □

The equations in $(*_D)$ and $(*_P)$ are often called *complementary slackness conditions*. These conditions can be exploited to solve for an optimal solution of the primal problem with the help of the dual problem, and conversely. Indeed, if we guess a solution to one problem, then we may solve for a solution of the dual using the complementary slackness conditions, and then check that our guess was correct. This is the essence of the *primal-dual* method. To present this method, first we need to take a closer look at the dual of a linear program already in standard form.

9.4 Duality for Linear Programs in Standard Form

Let (P) be a linear program in standard form, where $Ax = b$ for some $m \times n$ matrix of rank m and some objective function $x \mapsto cx$ (of course, $x \geq 0$). To obtain the dual of (P) we convert the equations $Ax = b$ to the following system of inequalities involving a $(2m) \times n$ matrix:

$$\begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix}.$$

Then if we denote the $2m$ dual variables by (y', y''), with $y', y'' \in (\mathbb{R}^m)^*$, the dual of the above program is

$$\begin{aligned} &\text{minimize} \quad y'b - y''b \\ &\text{subject to} \quad \begin{pmatrix} y' & y'' \end{pmatrix}\begin{pmatrix} A \\ -A \end{pmatrix} \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$, which is equivalent to

$$\begin{aligned} &\text{minimize} \quad (y' - y'')b \\ &\text{subject to} \quad (y' - y'')A \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$. If we write $y = y' - y''$, we find that the above linear program is equivalent to the following Linear Program (D):

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. Observe that y is *not required* to be nonnegative; it is arbitrary.

Next we would like to know what is the version of Theorem 9.8 for a linear program already in standard form. This is very simple.

Theorem 9.12. *Consider the Linear Program (P2) in standard form*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

and its Dual (D) given by

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. If the simplex algorithm applied to the Linear Program $(P2)$ terminates with an optimal solution (u^*, K^*), where u^* is a basic feasible solution and K^* is a basis for u^*, then $y^* = c_{K^*}A_{K^*}^{-1}$ is an optimal solution for (D) such that $cu^* = y^*b$. Furthermore, if we assume that the simplex algorithm is started with a basic feasible solution (u_0, K_0) where $K_0 = (n-m+1, \ldots, n)$ (the indices of the last m columns of A) and $A_{(n-m+1,\ldots,n)} = I_m$ (the last m columns of A constitute the identity matrix I_m), then the optimal solution $y^* = c_{K^*}A_{K^*}^{-1}$ for (D) is given in terms of the reduced costs by

$$y^* = c_{(n-m+1,\ldots,n)} - (\overline{c}_{K^*})_{(n-m+1,\ldots,n)},$$

and the $m \times m$ matrix consisting of last m columns and the last m rows of the final tableau is $A_{K^*}^{-1}$.

Proof. The proof of Theorem 9.8 applies with A instead of $\widehat{A}$, and we can show that

$$c_{K^*}A_{K^*}^{-1}A_{N^*} \geq c_{N^*},$$

and that $y^* = c_{K^*}A_{K^*}^{-1}$ satisfies, $cu^* = y^*b$, and

$$\begin{aligned} y^*A_{K^*} &= c_{K^*}A_{K^*}^{-1}A_{K^*} = c_{K^*}, \\ y^*A_{N^*} &= c_{K^*}A_{K^*}^{-1}A_{N^*} \geq c_{N^*}. \end{aligned}$$

Let P be the $n \times n$ permutation matrix defined so that

$$AP = \begin{pmatrix} A_{K^*} & A_{N^*} \end{pmatrix}.$$

Then we also have

$$cP = \begin{pmatrix} c_{K^*} & c_{N^*} \end{pmatrix},$$

and using the above equations and inequalities we obtain

$$y^* \begin{pmatrix} A_{K^*} & A_{N^*} \end{pmatrix} \geq \begin{pmatrix} c_{K^*} & c_{N^*} \end{pmatrix},$$

that is, $y^*AP \geq cP$, which is equivalent to

$$y^*A \geq c,$$

which shows that y^* is a feasible solution of (D) (remember, y^* is arbitrary so there is no need for the constraint $y^* \geq 0$).

The reduced costs are given by

$$(\overline{c}_{K^*})_i = c_i - c_{K^*}A_{K^*}^{-1}A^i,$$

and since for $j = n-m+1, \ldots, n$ the column A^j is the $(j+m-n)$th column of the identity matrix I_m, we have

$$(\overline{c}_{K^*})_j = c_j - (c_{K^*}A_{K^*}^{-1})_{j+m-n} \quad j = n-m+1, \ldots, n,$$

that is,

$$y^* = c_{(n-m+1,\ldots,n)} - (\overline{c}_{K^*})_{(n-m+1,\ldots,n)},$$

as claimed. Since the last m rows of the final tableau is obtained by multiplying $[u_0 \;\; A]$ by $A_{K^*}^{-1}$, and the last m columns of A constitute I_m, the last m rows and the last m columns of the final tableau constitute $A_{K^*}^{-1}$. □

Let us now take a look at the complementary slackness conditions of Theorem 9.11. If we go back to the version of (P) given by

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad \begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix} \text{ and } x \geq 0, \end{aligned}$$

and to the version of (D) given by

$$\begin{aligned} &\text{minimize} \quad y'b - y''b \\ &\text{subject to} \quad \begin{pmatrix} y' & y'' \end{pmatrix} \begin{pmatrix} A \\ -A \end{pmatrix} \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$, since the inequalities $Ax \leq b$ and $-Ax \leq -b$ together imply that $Ax = b$, we have equality for all these inequality constraints, and so the Conditions $(*_D)$ place no constraints at all on y' and y'', while the Conditions $(*_P)$ assert that

$$x_j = 0 \qquad \text{for all } j \text{ for which } \textstyle\sum_{i=1}^m (y'_i - y''_i)a_{ij} > c_j.$$

If we write $y = y' - y''$, the above conditions are equivalent to

$$x_j = 0 \qquad \text{for all } j \text{ for which } \sum_{i=1}^{m} y_i a_{ij} > c_j.$$

Thus we have the following version of Theorem 9.11.

Theorem 9.13. *(Equilibrium Theorem, Version 2) For any Linear Program $(P2)$ in standard form (with $Ax = b$ where A is an $m \times n$ matrix, $x \geq 0$, and objective function $x \mapsto cx$) and its Dual Linear Program (D), for any feasible solution x of (P) and any feasible solution y of (D), x and y are optimal solutions iff*

$$x_j = 0 \qquad \textit{for all } j \textit{ for which } \sum_{i=1}^{m} y_i a_{ij} > c_j. \tag{$*_P$}$$

Therefore, the slackness conditions applied to a Linear Program $(P2)$ in standard form and to its Dual (D) *only impose slackness conditions on the variables x_j of the primal problem.*

The above fact plays a crucial role in the primal-dual method.

9.5 The Dual Simplex Algorithm

Given a Linear Program $(P2)$ in standard form

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix of rank m, if no obvious feasible solution is available *but if* $c \leq 0$, rather than using the method for finding a feasible solution described in Section 8.2 we may use a method known as the dual simplex algorithm. This method uses basic solutions (u, K) where $Au = b$ and $u_j = 0$ for all $u_j \notin K$, but does not require $u \geq 0$, so u may not be feasible. However, $y = c_K A_K^{-1}$ is required to be feasible for the dual program

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^*)^m$. Since $c \leq 0$, observe that $y = 0_m^\top$ is a feasible solution of the dual.

If a basic solution u of $(P2)$ is found such that $u \geq 0$, then $cu = yb$ for $y = c_K A_K^{-1}$, and we have found an optimal solution u for $(P2)$ and y for (D). The dual simplex method makes progress by attempting to make negative components of u zero and by decreasing the objective function of the dual program.

The dual simplex method starts with a basic solution (u, K) of $Ax = b$ which is not feasible but for which $y = c_K A_K^{-1}$ is dual feasible. In many cases the original linear program is specified by a set of inequalities $Ax \leq b$ with some $b_i < 0$, so by adding slack variables it is

easy to find such basic solution u, and if in addition $c \leq 0$, then because the cost associated with slack variables is 0, we see that $y = 0$ is a feasible solution of the dual.

Given a basic solution (u, K) of $Ax = b$ (feasible or not), $y = c_K A_K^{-1}$ is dual feasible iff $c_K A_K^{-1} A \geq c$, and since $c_K A_K^{-1} A_K = c_K$, the inequality $c_K A_K^{-1} A \geq c$ is equivalent to $c_K A_K^{-1} A_N \geq c_N$, that is,

$$c_N - c_K A_K^{-1} A_N \leq 0, \tag{$*_1$}$$

where $N = \{1, \ldots, n\} - K$. Equation $(*_1)$ is equivalent to

$$c_j - c_K \gamma_K^j \leq 0 \quad \text{for all } j \in N, \tag{$*_2$}$$

where $\gamma_K^j = A_K^{-1} A^j$. Recall that the notation $\overline{c}_j$ is used to denote $c_j - c_K \gamma_K^j$, which is called the *reduced cost* of the variable x_j.

As in the simplex algorithm we need to decide which column A^k leaves the basis K and which column A^j enters the new basis K^+, in such a way that $y^+ = c_{K^+} A_{K^+}^{-1}$ is a feasible solution of (D), that is, $c_{N^+} - c_{K^+} A_{K^+}^{-1} A_{N^+} \leq 0$, where $N^+ = \{1, \ldots, n\} - K^+$. We use Proposition 8.2 to decide wich column k^- should leave the basis.

Suppose (u, K) is a solution of $Ax = b$ for which $y = c_K A_K^{-1}$ is dual feasible.

Case (A). If $u \geq 0$, then u is an optimal solution of $(P2)$.

Case (B). There is some $k \in K$ such that $u_k < 0$. In this case pick some $k^- \in K$ such that $u_{k^-} < 0$ (according to some pivot rule).

Case (B1). Suppose that $\gamma_{k^-}^j \geq 0$ for all $j \notin K$ (in fact, for all j, since $\gamma_{k^-}^j \in \{0, 1\}$ for all $j \in K$). If so, we we claim that $(P2)$ is not feasible.

Indeed, let v be some basic feasible solution. We have $v \geq 0$ and $Av = b$, that is,

$$\sum_{j=1}^{n} v_j A^j = b,$$

so by multiplying both sides by A_K^{-1} and using the fact that by definition $\gamma_K^j = A_K^{-1} A^j$, we obtain

$$\sum_{j=1}^{n} v_j \gamma_K^j = A_K^{-1} b = u_K.$$

But recall that by hypothesis $u_{k^-} < 0$, yet $v_j \geq 0$ and $\gamma_{k^-}^j \geq 0$ for all j, so the component of index k^- is zero or positive on the left, and negative on the right, a contradiction. Therefore, $(P2)$ is indeed not feasible.

Case (B2). We have $\gamma_{k^-}^j < 0$ for some j.

We pick the column A^j entering the basis among those for which $\gamma_{k^-}^j < 0$. Since we assumed that $c_j - c_K \gamma_K^j \leq 0$ for all $j \in N$ by $(*_2)$, consider

$$\mu^+ = \max\left\{ -\frac{c_j - c_K \gamma_K^j}{\gamma_{k^-}^j} \;\middle|\; \gamma_{k^-}^j < 0,\ j \in N \right\} = \max\left\{ -\frac{\overline{c}_j}{\gamma_{k^-}^j} \;\middle|\; \gamma_{k^-}^j < 0,\ j \in N \right\} \leq 0,$$

and the set

$$N(\mu^+) = \left\{ j \in N \;\middle|\; -\frac{\overline{c}_j}{\gamma^j_{k^-}} = \mu^+ \right\}.$$

We pick some index $j^+ \in N(\mu^+)$ as the index of the column entering the basis (using some pivot rule).

Recall that by hypothesis $c_i - c_K\gamma^i_K \leq 0$ for all $j \notin K$ and $c_i - c_K\gamma^i_K = 0$ for all $i \in K$. Since $\gamma^{j^+}_{k^-} < 0$, for any index i such that $\gamma^i_{k^-} \geq 0$, we have $-\gamma^i_{k^-}/\gamma^{j^+}_{k^-} \geq 0$, and since by Proposition 8.2

$$c_i - c_{K^+}\gamma^i_{K^+} = c_i - c_K\gamma^i_K - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(c_{j^+} - c_K\gamma^{j^+}_K),$$

we have $c_i - c_{K^+}\gamma^i_{K^+} \leq 0$. For any index i such that $\gamma^i_{k^-} < 0$, by the choice of $j^+ \in K^*$,

$$-\frac{c_i - c_K\gamma^i_K}{\gamma^i_{k^-}} \leq -\frac{c_{j^+} - c_K\gamma^{j^+}_K}{\gamma^{j^+}_{k^-}},$$

so

$$c_i - c_K\gamma^i_K - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(c_{j^+} - c_K\gamma^{j^+}_K) \leq 0,$$

and again, $c_i - c_{K^+}\gamma^i_{K^+} \leq 0$. Therefore, if we let $K^+ = (K - \{k^-\}) \cup \{j^+\}$, then $y^+ = c_{K^+}A^{-1}_{K^+}$ is dual feasible. As in the simplex algorithm, θ^+ is given by

$$\theta^+ = u_{k^-}/\gamma^{j^+}_{k^-} \geq 0,$$

and u^+ is also computed as in the simplex algorithm by

$$u^+_i = \begin{cases} u_i - \theta^{j^+}\gamma^{j^+}_i & \text{if } i \in K \\ \theta^{j^+} & \text{if } i = j^+ \\ 0 & \text{if } i \notin K \cup \{j^+\} \end{cases}.$$

The change in the objective function of the primal and dual program (which is the same, since $u_K = A^{-1}_K b$ and $y = c_K A^{-1}_K$ is chosen such that $cu = c_K u_K = yb$) is the same as in the simplex algorithm, namely

$$\theta^+\left(c^{j^+} - c_K\gamma^{j^+}_K\right).$$

We have $\theta^+ > 0$ and $c^{j^+} - c_K\gamma^{j^+}_K \leq 0$, so if $c^{j^+} - c_K\gamma^{j^+}_K < 0$, then the objective function of the dual program decreases strictly.

Case (B3). $\mu^+ = 0$.

The possibity that $\mu^+ = 0$, that is, $c^{j^+} - c_K\gamma^{j^+}_K = 0$, may arise. In this case, the objective function doesn't change. This is a case of degeneracy similar to the degeneracy that arises in the simplex algorithm. We still pick $j^+ \in N(\mu^+)$, but we need a pivot rule that prevents

cycling. Such rules exist; see Bertsimas and Tsitsiklis [10] (Section 4.5) and Papadimitriou and Steiglitz [47] (Section 3.6).

The reader surely noticed that the dual simplex algorithm is very similar to the simplex algorithm, except that the simplex algorithm preserves the property that (u, K) is (primal) feasible, whereas the dual simplex algorithm preserves the property that $y = c_K A_K^{-1}$ is dual feasible. One might then wonder whether the dual simplex algorithm is equivalent to the simplex algorithm applied to the dual problem. This is indeed the case, there is a one-to-one correspondence between the dual simplex algorithm and the simplex algorithm applied to the dual problem in maximization form. This correspondence is described in Papadimitriou and Steiglitz [47] (Section 3.7).

The comparison between the simplex algorithm and the dual simplex algorithm is best illustrated if we use a description of these methods in terms of *(full) tableaux*.

Recall that a *(full) tableau* is an $(m+1) \times (n+1)$ matrix organized as follows:

$-c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ_1^1	$\cdots$	γ_1^j	$\cdots$	γ_1^n
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ_m^1	$\cdots$	γ_m^j	$\cdots$	γ_m^n

The top row contains the current value of the objective function and the reduced costs, the first column except for its top entry contain the components of the current basic solution u_K, and the remaining columns except for their top entry contain the vectors γ_K^j. Observe that the γ_K^j corresponding to indices j in K constitute a permutation of the identity matrix I_m. A tableau together with the new basis $K^+ = (K - \{k^-\}) \cup \{j^+\}$ contains all the data needed to compute the new u_{K^+}, the new $\gamma_{K^+}^j$, and the new reduced costs $\overline{c}_i - (\gamma_{k^-}^i / \gamma_{k^-}^{j^+}) \overline{c}_{j^+}$.

When executing the simplex algorithm, we have $u_k \geq 0$ for all $k \in K$ (and $u_j = 0$ for all $j \notin K$), and the incoming column j^+ is determined by picking one of the column indices such that $\overline{c}_j > 0$. Then the index k^- of the leaving column is determined by looking at the minimum of the ratios $u_k / \gamma_k^{j^+}$ for which $\gamma_k^{j^+} > 0$ (along column j^+).

On the other hand, when executing the dual simplex algorithm, we have $\overline{c}_j \leq 0$ for all $j \notin K$ (and $\overline{c}_k = 0$ for all $k \in K$), and the outgoing column k^- is determined by picking one of the row indices such that $u_k < 0$. The index j^+ of the incoming column is determined by looking at the maximum of the ratios $-\overline{c}_j / \gamma_{k^-}^j$ for which $\gamma_{k^-}^j < 0$ (along row k^-).

More details about the comparison between the simplex algorithm and the dual simplex algorithm can be found in Bertsimas and Tsitsiklis [10] and Papadimitriou and Steiglitz [47].

Here is an example of the the dual simplex method.

Example 9.2. Consider the following linear program in standard form:

Maximize $-4x_1 - 2x_2 - x_3$

$$\text{subject to} \quad \begin{pmatrix} -1 & -1 & 2 & 1 & 0 & 0 \\ -4 & -2 & 1 & 0 & 1 & 0 \\ 1 & 1 & -4 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} -3 \\ -4 \\ 2 \end{pmatrix} \quad \text{and} \quad x_1, x_2, x_3, x_4, x_5, x_6 \geq 0.$$

We initialize the dual simplex procedure with (u, K) where $u = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -3 \\ -4 \\ 2 \end{pmatrix}$ and $K = (4, 5, 6)$.

The initial tableau, before explicitly calculating the reduced cost, is

0	$\overline{c}_1$	$\overline{c}_2$	$\overline{c}_3$	$\overline{c}_4$	$\overline{c}_5$	$\overline{c}_6$
$u_4 = -3$	-1	-1	2	1	0	0
$u_5 = -4$	-4	-2	1	0	1	0
$u_6 = 2$	1	1	-4	0	0	1

.

Since u has negative coordinates, Case (B) applies, and we will set $k^- = 4$. We must now determine whether Case (B1) or Case (B2) applies. This determination is accomplished by scanning the first three columns in the tableau and observing each column has a negative entry. Thus Case (B2) is applicable, and we need to determine the reduced costs. Observe that $c = (-4, -2, -1, 0, 0, 0)$, which in turn implies $c_{(4,5,6)} = (0, 0, 0)$. Equation $(*_2)$ implies that the nonzero reduced costs are

$$\overline{c}_1 = c_1 - c_{(4,5,6)} \begin{pmatrix} -1 \\ -4 \\ 1 \end{pmatrix} = -4$$

$$\overline{c}_2 = c_2 - c_{(4,5,6)} \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix} = -2$$

$$\overline{c}_3 = c_3 - c_{(4,5,6)} \begin{pmatrix} -2 \\ 1 \\ 4 \end{pmatrix} = -1,$$

and our tableau becomes

0	-4	-2	-1	0	0	0
$u_4 = -3$	-1	-1	2	1	0	0
$u_5 = -4$	-4	-2	1	0	1	0
$u_6 = 2$	1	1	-4	0	0	1

.

Since $k^- = 4$, our pivot row is the first row of the tableau. To determine candidates for j^+, we scan this row, locate negative entries and compute

$$\mu^+ = \max\left\{ -\frac{\overline{c}_j}{\gamma_4^j} \;\middle|\; \gamma_4^j < 0,\ j \in \{1,2,3\} \right\} = \max\left\{ \frac{-2}{1}, \frac{-4}{1} \right\} = -2.$$

Since μ^+ occurs when $j = 2$, we set $j^+ = 2$. Our new basis is $K^+ = (2,5,6)$. We must normalize the first row of the tableau, namely multiply by -1, then add twice this normalized row to the second row, and subtract the normalized row from the third row to obtain the updated tableau.

0	-4	-2	-1	0	0	0
$u_2 = 3$	1	(1)	-2	-1	0	0
$u_5 = 2$	-2	0	-3	-2	1	0
$u_6 = -1$	0	0	-2	1	0	1

It remains to update the reduced costs and the value of the objective function by adding twice the normalized row to the top row.

6	-2	0	-5	-2	0	0
$u_2 = 3$	1	1	-2	-1	0	0
$u_5 = 2$	-2	0	-3	-2	1	0
$u_6 = -1$	0	0	(−2)	1	0	1

We now repeat the procedure of Case (B2) and set $k^- = 6$ (since this is the only negative entry of u^+). Our pivot row is now the third row of the updated tableau, and the new μ^+ becomes

$$\mu^+ = \max\left\{ -\frac{\overline{c}_j}{\gamma_6^j} \;\middle|\; \gamma_6^j < 0,\ j \in \{1,3,4\} \right\} = \max\left\{ \frac{-5}{2} \right\} = -\frac{5}{2},$$

which implies that $j^+ = 3$. Hence the new basis is $K^+ = (2,5,3)$, and we update the tableau by taking $-\frac{1}{2}$ of Row 3, adding twice the normalized Row 3 to Row 1, and adding three times the normalized Row 3 to Row 2.

6	-2	0	-5	-2	0	0
$u_2 = 4$	1	1	0	-2	0	-1
$u_5 = 7/2$	-2	0	0	$-7/2$	1	$-3/2$
$u_3 = 1/2$	0	0	(1)	$-1/2$	0	$-1/2$

It remains to update the objective function and the reduced costs by adding five times the normalized row to the top row.

17/2	-2	0	0	$-9/2$	0	$-5/2$
$u_2 = 4$	1	1	0	-2	0	-1
$u_5 = 7/2$	-2	0	0	$-\frac{7}{2}$	1	$-3/2$
$u_3 = 1/2$	0	0	(1)	$-1/2$	0	$-1/2$

Since u^+ has no negative entries, the dual simplex method terminates and objective function $-4x_1 - 2x_2 - x_3$ is maximized with $-\frac{17}{2}$ at $(0, 4, \frac{1}{2})$. See Figure 9.4.

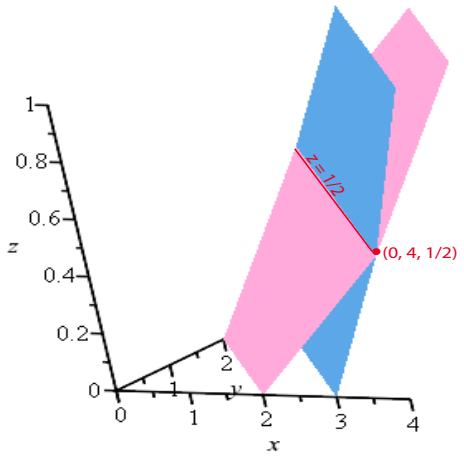


Figure 9.4: The objective function $-4x_1 - 2x_2 - x_3$ is maximized at the intersection between the blue plane $-x_1 - x_2 + 2x_3 = -3$ and the pink plane $x_1 + x_2 - 4x_3 = 2$.

9.6 The Primal-Dual Algorithm

Let $(P2)$ be a linear program in standard form

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix of rank m, and (D) be its dual given by

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$.

First we may assume that $b \geq 0$ by changing every equation $\sum_{j=1}^n a_{ij}x_j = b_i$ with $b_i < 0$ to $\sum_{j=1}^n -a_{ij}x_j = -b_i$. If we happen to have some feasible solution y of the dual program (D), we know from Theorem 9.13 that a feasible solution x of $(P2)$ is an optimal solution iff the equations in $(*_P)$ hold. If we denote by J the subset of $\{1, \ldots, n\}$ for which the equalities

$$yA^j = c_j$$

hold, then by Theorem 9.13 a feasible solution x of $(P2)$ is an optimal solution iff

$$x_j = 0 \quad \text{for all } j \notin J.$$

Let $|J| = p$ and $N = \{1, \ldots, n\} - J$. The above suggests looking for $x \in \mathbb{R}^n$ such that

$$\sum_{j \in J} x_j A^j = b$$
$$x_j \geq 0 \quad \text{for all } j \in J$$
$$x_j = 0 \quad \text{for all } j \notin J,$$

or equivalently

$$A_J x_J = b, \quad x_J \geq 0, \tag{$*_1$}$$

and

$$x_N = 0_{n-p}.$$

To search for such an x, we just need to look for a feasible x_J, and for this we can use the *Restricted Primal* linear program (RP) defined as follows:

$$\begin{aligned} &\text{maximize} && -(\xi_1 + \cdots + \xi_m) \\ &\text{subject to} && \begin{pmatrix} A_J & I_m \end{pmatrix} \begin{pmatrix} x_J \\ \xi \end{pmatrix} = b \text{ and } x, \xi \geq 0. \end{aligned}$$

Since by hypothesis $b \geq 0$ and the objective function is bounded above by 0, this linear program has an optimal solution (x_J^*, ξ^*).

If $\xi^* = 0$, then the vector $u^* \in \mathbb{R}^n$ given by $u_J^* = x_J^*$ and $u_N^* = 0_{n-p}$ is an optimal solution of (P).

Otherwise, $\xi^* > 0$ and we have failed to solve $(*_1)$. However we may try to use ξ^* to improve y. For this consider the Dual (DRP) of (RP):

$$\begin{aligned} &\text{minimize} && zb \\ &\text{subject to} && zA_J \geq 0 \\ & && z \geq -\mathbf{1}_m^\top. \end{aligned}$$

Observe that the Program (DRP) has the same objective function as the original Dual Program (D). We know by Theorem 9.12 that the optimal solution (x_J^*, ξ^*) of (RP) yields an optimal solution z^* of (DRP) such that

$$z^* b = -(\xi_1^* + \cdots + \xi_m^*) < 0.$$

In fact, if K^* is the basis associated with (x_J^*, ξ^*) and if we write

$$\widehat{A} = \begin{pmatrix} A_J & I_m \end{pmatrix}$$

and $\widehat{c} = [0_p^\top \quad -\mathbf{1}^\top]$, then by Theorem 9.12 we have

$$z^* = \widehat{c}_{K^*} \widehat{A}_{K^*}^{-1} = -\mathbf{1}_m^\top - (\overline{c}_{K^*})_{(p+1,\ldots,p+m)},$$

where $(\overline{c}_{K^*})_{(p+1,\ldots,p+m)}$ denotes the row vector of reduced costs in the final tableau corresponding to the last m columns.

If we write

$$y(\theta) = y + \theta z^*,$$

then the new value of the objective function of (D) is

$$y(\theta)b = yb + \theta z^* b, \tag{$*_2$}$$

and since $z^*b < 0$, we have a chance of improving the objective function of (D), that is, decreasing its value for $\theta > 0$ small enough if $y(\theta)$ is feasible for (D). This will be the case iff $y(\theta)A \geq c$ iff

$$yA + \theta z^* A \geq c. \tag{$*_3$}$$

Now since y is a feasible solution of (D) we have $yA \geq c$, so if $z^*A \geq 0$, then $(*_3)$ is satisfied and $y(\theta)$ is a solution of (D) for all $\theta > 0$, which means that (D) is unbounded. But this implies that (P) is not feasible.

Let us take a closer look at the inequalities $z^*A \geq 0$. For $j \in J$, since z^* is an optimal solution of (DRP), we know that $z^*A_J \geq 0$, so if $z^*A^j \geq 0$ for all $j \in N$, then $(P2)$ is not feasible.

Otherwise, there is some $j \in N = \{1, \ldots, n\} - J$ such that

$$z^*A^j < 0,$$

and then since by the definition of N we have $yA^j > c_j$ for all $j \in N$, if we pick θ such that

$$0 < \theta \leq \frac{yA^j - c_j}{-z^*A^j} \quad j \in N,\ z^*A^j < 0,$$

then we decrease the objective function $y(\theta)b = yb + \theta z^* b$ of (D) (since $z^*b < 0$). Therefore we pick the best θ, namely

$$\theta^+ = \min\left\{ \frac{yA^j - c_j}{-z^*A^j} \;\middle|\; j \notin J,\ z^*A^j < 0 \right\} > 0. \tag{$*_4$}$$

Next we update y to $y^+ = y(\theta^+) = y + \theta^+ z^*$, we create the new restricted primal with the new subset

$$J^+ = \{j \in \{1, \ldots, n\} \mid y^+A^j = c_j\},$$

and repeat the process.

Here are the steps of the primal-dual algorithm.

Step 1. Find some feasible solution y of the Dual Program (D). We will show later that this is always possible.

Step 2. Compute

$$J^+ = \{j \in \{1, \ldots, n\} \mid yA^j = c_j\}.$$

Step 3. Set $J = J^+$ and solve the Problem (RP) using the simplex algorithm, starting from the optimal solution determined during the previous round, obtaining the optimal solution (x_J^*, ξ^*) with the basis K^*.

Step 4.

If $\xi^* = 0$, then stop with an optimal solution u^* for (P) such that $u_J^* = x_J^*$ and the other components of u^* are zero.

Else let

$$z^* = -\mathbf{1}_m^\top - (\overline{c}_{K^*})_{(p+1,\ldots,p+m)},$$

be the optimal solution of (DRP) corresponding to (x_J^*, ξ^*) and the basis K^*.

If $z^*A^j \geq 0$ for all $j \notin J$, then stop; the Program (P) has no feasible solution.

Else compute

$$\theta^+ = \min\left\{ -\frac{yA^j - c_j}{z^*A^j} \;\middle|\; j \notin J,\ z^*A^j < 0 \right\}, \qquad y^+ = y + \theta^+ z^*,$$

and

$$J^+ = \{j \in \{1, \ldots, n\} \mid y^+A^j = c_j\}.$$

Go back to Step 3.

The following proposition shows that at each iteration we can start the Program (RP) with the optimal solution obtained at the previous iteration.

Proposition 9.14. *Every $j \in J$ such that A^j is in the basis of the optimal solution ξ^* belongs to the next index set J^+.*

Proof. Such an index $j \in J$ correspond to a variable ξ_j such that $\xi_j > 0$, so by complementary slackness, the constraint $z^*A^j \geq 0$ of the Dual Program (DRP) must be an equality, that is, $z^*A^j = 0$. But then we have

$$y^+A^j = yA^j + \theta^+ z^*A^j = c_j,$$

which shows that $j \in J^+$. □

If (u^*, ξ^*) with the basis K^* is the optimal solution of the Program (RP), Proposition 9.14 together with the last property of Theorem 9.12 allows us to restart the (RP) in Step 3 with $(u^*, \xi^*)_{K^*}$ as initial solution (with basis K^*). For every $j \in J - J^+$, column j is deleted, and for every $j \in J^+ - J$, the new column A^j is computed by multiplying $\widehat{A}_{K^*}^{-1}$ and A^j, but $\widehat{A}_{K^*}^{-1}$ is the matrix $\Gamma^*[1{:}m; p+1{:}p+m]$ consisting of the last m columns of Γ^* in the final

tableau, and the new reduced $\overline{c}_j$ is given by $c_j - z^* A^j$. Reusing the optimal solution of the previous (RP) may improve efficiency significantly.

Another crucial observation is that for any index $j_0 \in N$ such that $\theta^+ = (yA^{j_0} - c_{j_0})/(-z^* A^{j_0})$, we have

$$y^+ A_{j_0} = yA_{j_0} + \theta^+ z^* A^{j_0} = c_{j_0},$$

and so $j_0 \in J^+$. This fact that be used to ensure that the primal-dual algorithm terminates in a finite number of steps (using a pivot rule that prevents cycling); see Papadimitriou and Steiglitz [47] (Theorem 5.4).

It remains to discuss how to pick some initial feasible solution y of the Dual Program (D). If $c_j \leq 0$ for $j = 1, \ldots, n$, then we can pick $y = 0$. If we are dealing with a minimization problem, the weight c_j are often nonnegative, so from the point of view of maximization we will have $-c_j \leq 0$ for all j, and we will be able to use $y = 0$ as a starting point.

Going back to our primal problem in maximization form and its dual in minimization form, we still need to deal with the situation where $c_j > 0$ for some j, in which case there may not be any obvious y feasible for (D). Preferably we would like to find such a y very cheaply.

There is a trick to deal with this situation. We pick some very large positive number M and add to the set of equations $Ax = b$ the new equation

$$x_1 + \cdots + x_n + x_{n+1} = M,$$

with the new variable x_{n+1} constrained to be nonnegative. If the Program (P) has a feasible solution, such an M exists. In fact it can shown that for any basic feasible solution $u = (u_1, \ldots, u_n)$, each $|u_i|$ is bounded by some expression depending only on A and b; see Papadimitriou and Steiglitz [47] (Lemma 2.1). The proof is not difficult and relies on the fact that the inverse of a matrix can be expressed in terms of certain determinants (the adjugates). Unfortunately, this bound contains $m!$ as a factor, which makes it quite impractical.

Having added the new equation above, we obtain the new set of equations

$$\begin{pmatrix} A & 0_n \\ \mathbf{1}_n^\top & 1 \end{pmatrix} \begin{pmatrix} x \\ x_{n+1} \end{pmatrix} = \begin{pmatrix} b \\ M \end{pmatrix},$$

with $x \geq 0, x_{n+1} \geq 0$, and the new objective function given by

$$\begin{pmatrix} c & 0 \end{pmatrix} \begin{pmatrix} x \\ x_{n+1} \end{pmatrix} = cx.$$

The dual of the above linear program is

$$\begin{array}{ll} \text{minimize} & yb + y_{m+1} M \\ \text{subject to} & yA^j + y_{m+1} \geq c_j \quad j = 1, \ldots, n \\ & y_{m+1} \geq 0. \end{array}$$

If $c_j > 0$ for some j, observe that the linear form $\widetilde{y}$ given by

$$\widetilde{y}_i = \begin{cases} 0 & \text{if } 1 \leq i \leq m \\ \max_{1 \leq j \leq n}\{c_j\} > 0 & \end{cases}$$

is a feasible solution of the new dual program. In practice, we can choose M to be a number close to the largest integer representable on the computer being used.

Here is an example of the primal-dual algorithm given in the Math 588 class notes of T. Molla [44].

Example 9.3. Consider the following linear program in standard form:

$$\begin{aligned} &\text{Maximize} && -x_1 - 3x_2 - 3x_3 - x_4 \\ &\text{subject to} && \begin{pmatrix} 3 & 4 & -3 & 1 \\ 3 & -2 & 6 & -1 \\ 6 & 4 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \text{ and } x_1, x_2, x_3, x_4 \geq 0. \end{aligned}$$

The associated Dual Program (D) is

$$\begin{aligned} &\text{Minimize} && 2y_1 + y_2 + 4y_3 \\ &\text{subject to} && \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix} \begin{pmatrix} 3 & 4 & -3 & 1 \\ 3 & -2 & 6 & -1 \\ 6 & 4 & 0 & 1 \end{pmatrix} \geq \begin{pmatrix} -1 & -3 & -3 & -1 \end{pmatrix}. \end{aligned}$$

We initialize the primal-dual algorithm with the dual feasible point $y = (-1/3 \;\; 0 \;\; 0)$. Observe that only the first inequality of (D) is actually an equality, and hence $J = \{1\}$. We form the Restricted Primal Program $(RP1)$

$$\begin{aligned} &\text{Maximize} && -(\xi_1 + \xi_2 + \xi_3) \\ &\text{subject to} && \begin{pmatrix} 3 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ 6 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \text{ and } x_1, \xi_1, \xi_2, \xi_3 \geq 0. \end{aligned}$$

We now solve $(RP1)$ via the simplex algorithm. The initial tableau with $K = (2, 3, 4)$ and $J = \{1\}$ is

	x_1	ξ_1	ξ_2	ξ_3
7	12	0	0	0
$\xi_1 = 2$	3	1	0	0
$\xi_2 = 1$	(3)	0	1	0
$\xi_3 = 4$	6	0	0	1

.

For $(RP1)$, $\hat{c} = (0, -1, -1, -1)$, $(x_1, \xi_1, \xi_2, \xi_3) = (0, 2, 1, 4)$, and the nonzero reduced cost is given by

$$0 - (-1 \;\; -1 \;\; -1) \begin{pmatrix} 3 \\ 3 \\ 6 \end{pmatrix} = 12.$$

Since there is only one nonzero reduced cost, we must set $j^+ = 1$. Since $\min\{\xi_1/3, \xi_2/3, \xi_3/6\} = 1/3$, we see that $k^- = 3$ and $K = (2, 1, 4)$. Hence we pivot through the red circled 3 (namely we divide row 2 by 3, and then subtract $3\times$ (row 2) from row 1, $6\times$ (row 2) from row 3, and $12\times$ (row 2) from row 0), to obtain the tableau

	x_1	ξ_1	ξ_2	ξ_3
3	0	0	-4	0
$\xi_1 = 1$	0	1	-1	0
$x_1 = 1/3$	1	0	$1/3$	0
$\xi_3 = 2$	0	0	-2	1

.

At this stage the simplex algorithm for $(RP1)$ terminates since there are no positive reduced costs. Since the upper left corner of the final tableau is not zero, we proceed with Step 4 of the primal dual algorithm and compute

$$z^* = (-1 \;\; -1 \;\; -1) - (0 \;\; -4 \;\; 0) = (-1 \;\; 3 \;\; -1),$$

$$yA^2 - c_2 = (-1/3 \;\; 0 \;\; 0) \begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} + 3 = \frac{5}{3}, \qquad z^*A^2 = -(-1 \;\; 3 \;\; -1) \begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} = 14,$$

$$yA^4 - c_4 = (-1/3 \;\; 0 \;\; 0) \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + 1 = \frac{2}{3}, \qquad z^*A^4 = -(-1 \;\; 3 \;\; -1) \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 5,$$

so

$$\theta^+ = \min\left\{ \frac{5}{42}, \frac{2}{15} \right\} = \frac{5}{42},$$

and we conclude that the new feasible solution for (D) is

$$y^+ = (-1/3 \;\; 0 \;\; 0) + \frac{5}{42}(-1 \;\; 3 \;\; -1) = (-19/42 \;\; 5/14 \;\; -5/42).$$

When we substitute y^+ into (D), we discover that the first two constraints are equalities, and that the new J is $J = \{1, 2\}$. The new Reduced Primal $(RP2)$ is

$$\begin{aligned} &\text{Maximize} \quad -(\xi_1 + \xi_2 + \xi_3) \\ &\text{subject to} \quad \begin{pmatrix} 3 & 4 & 1 & 0 & 0 \\ 3 & -2 & 0 & 1 & 0 \\ 6 & 4 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \text{ and } x_1, x_2, \xi_1, \xi_2, \xi_3 \geq 0. \end{aligned}$$

Once again, we solve $(RP2)$ via the simplex algorithm, where $\hat{c} = (0, 0, -1, -1, -1)$, $(x_1, x_2, \xi_1, \xi_2, \xi_3) = (1/3, 0, 1, 0, 2)$ and $K = (3, 1, 5)$. The initial tableau is obtained from the final tableau of the previous $(RP1)$ by adding a column corresponding the the variable x_2, namely

$$\widehat{A}_K^{-1} A^2 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1/3 & 0 \\ 0 & -2 & 1 \end{pmatrix} \begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} = \begin{pmatrix} 6 \\ -2/3 \\ 8 \end{pmatrix},$$

with

$$\bar{c}_2 = c_2 - z^* A^2 = 0 - \begin{pmatrix} -1 & 3 & -1 \end{pmatrix} \begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} = 14,$$

and we get

	x_1	x_2	ξ_1	ξ_2	ξ_3
3	0	14	0	-4	0
$\xi_1 = 1$	0	(6)	1	-1	0
$x_1 = 1/3$	1	$-2/3$	0	$1/3$	0
$\xi_3 = 2$	0	8	0	-2	1

.

Note that $j^+ = 2$ since the only positive reduced cost occurs in column 2. Also observe that since $\min\{\xi_1/6, \xi_3/8\} = \xi_1/6 = 1/6$, we set $k^- = 3$, $K = (2, 1, 5)$ and pivot along the red 6 to obtain the tableau

	x_1	x_2	ξ_1	ξ_2	ξ_3
2/3	0	0	$-7/3$	$-5/3$	0
$x_2 = 1/6$	0	1	$1/6$	$-1/6$	0
$x_1 = 4/9$	1	0	$1/9$	$2/9$	0
$\xi_3 = 2/3$	0	0	$-4/3$	$-2/3$	1

.

Since the reduced costs are either zero or negative the simplex algorithm terminates, and we compute

$$z^* = (-1 \ \ -1 \ \ -1) - (-7/3 \ \ -5/3 \ \ 0) = (4/3 \ \ 2/3 \ \ -1),$$

$$y^+ A^4 - c_4 = (-19/42 \ \ 5/14 \ \ -5/42) \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + 1 = 1/14,$$

$$z^* A^4 = -(4/3 \ \ 2/3 \ \ -1) \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 1/3,$$

so

$$\theta^+ = \frac{3}{14},$$

$$y^+ = (-19/42 \ \ 5/14 \ \ -5/42) + \frac{5}{14}(4/3 \ \ 2/3 \ \ -1) = (-1/6 \ \ 1/2 \ \ -1/3).$$

When we plug y^+ into (D), we discover that the first, second, and fourth constraints are equalities, which implies $J = \{1, 2, 4\}$. Hence the new Restricted Primal $(RP3)$ is

$$\begin{aligned} &\text{Maximize} \quad -(\xi_1 + \xi_2 + \xi_3) \\ &\text{subject to} \quad \begin{pmatrix} 3 & 4 & 1 & 1 & 0 & 0 \\ 3 & -2 & -1 & 0 & 1 & 0 \\ 6 & 4 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_4 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \quad \text{and} \quad x_1, x_2, x_4, \xi_1, \xi_2, \xi_3 \geq 0. \end{aligned}$$

The initial tableau for $(RP3)$, with $\hat{c} = (0, 0, 0, -1, -1, -1)$, $(x_1, x_2, x_4, \xi_1, \xi_2, \xi_3) = (4/9, 1/6, 0, 0, 0, 2/3)$ and $K = (2, 1, 6)$, is obtained from the final tableau of the previous $(RP2)$ by adding a column corresponding the the variable x_4, namely

$$\widehat{A}_K^{-1} A^4 = \begin{pmatrix} 1/6 & -1/6 & 0 \\ 1/9 & 2/9 & 0 \\ -4/3 & -2/3 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/3 \\ -1/9 \\ 1/3 \end{pmatrix},$$

with

$$\overline{c}_4 = c_4 - z^* A^4 = 0 - \begin{pmatrix} 4/3 & 2/3 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 1/3,$$

and we get

	x_1	x_2	x_4	ξ_1	ξ_2	ξ_3
$2/3$	0	0	$1/3$	$-7/3$	$-5/3$	0
$x_2 = 1/6$	0	1	$1/3$	$1/6$	$-1/6$	0
$x_1 = 4/9$	1	0	$-1/9$	$1/9$	$2/9$	0
$\xi_3 = 2/3$	0	0	$1/3$	$-4/3$	$-2/3$	1

.

Since the only positive reduced cost occurs in column 3, we set $j^+ = 3$. Furthermore since $\min\{x_2/(1/3), \xi_3/(1/3)\} = x_2/(1/3) = 1/2$, we let $k^- = 2$, $K = (3, 1, 6)$, and pivot around the red circled $1/3$ to obtain

	x_1	x_2	x_4	ξ_1	ξ_2	ξ_3
$1/2$	0	-1	0	$-5/2$	$-3/2$	0
$x_4 = 1/2$	0	3	1	$1/2$	$-1/2$	0
$x_1 = 1/2$	1	$1/3$	0	$1/6$	$1/6$	0
$\xi_3 = 1/2$	0	-1	0	$-3/2$	$-1/2$	1

.

At this stage there are no positive reduced costs, and we must compute

$$z^* = (-1 \quad -1 \quad -1) - (-5/2 \quad -3/2 \quad 0) = (3/2 \quad 1/2 \quad -1),$$

$$y^+A^3 - c_3 = (-1/6 \;\; 1/2 \;\; -1/3)\begin{pmatrix}-3\\6\\0\end{pmatrix} + 3 = 13/2,$$

$$z^*A^3 = -(3/2 \;\; 1/2 \;\; -1)\begin{pmatrix}-3\\6\\0\end{pmatrix} = 3/2,$$

so

$$\theta^+ = \frac{13}{3},$$
$$y^+ = (-1/6 \;\; 1/2 \;\; -1/3) + \frac{13}{3}(3/2 \;\; 1/2 \;\; -1) = (19/3 \;\; 8/3 \;\; -14/3).$$

We plug y^+ into (D) and discover that the first, third, and fourth constraints are equalities. Thus, $J = \{1, 3, 4\}$ and the Restricted Primal $(RP4)$ is

$$\text{Maximize} \quad -(\xi_1 + \xi_2 + \xi_3)$$
$$\text{subject to} \quad \begin{pmatrix}3 & -3 & 1 & 1 & 0 & 0\\3 & 6 & -1 & 0 & 1 & 0\\6 & 0 & 1 & 0 & 0 & 1\end{pmatrix}\begin{pmatrix}x_1\\x_3\\x_4\\\xi_1\\\xi_2\\\xi_3\end{pmatrix} = \begin{pmatrix}2\\1\\4\end{pmatrix} \quad \text{and} \quad x_1, x_3, x_4, \xi_1, \xi_2, \xi_3 \geq 0.$$

The initial tableau for $(RP4)$, with $\hat{c} = (0, 0, 0, -1, -1, -1)$, $(x_1, x_3, x_4, \xi_1, \xi_2, \xi_3) = (1/2, 0, 1/2, 0, 0, 1/2)$ and $K = (3, 1, 6)$ is obtained from the final tableau of the previous $(RP3)$ by replacing the column corresponding to the variable x_2 by a column corresponding to the variable x_3, namely

$$\widehat{A}_K^{-1}A^3 = \begin{pmatrix}1/2 & -1/2 & 0\\1/6 & 1/6 & 0\\-3/2 & -1/2 & 1\end{pmatrix}\begin{pmatrix}-3\\6\\0\end{pmatrix} = \begin{pmatrix}-9/2\\1/2\\3/2\end{pmatrix},$$

with

$$\overline{c}_3 = c_3 - z^*A^3 = 0 - \begin{pmatrix}3/2 & 1/2 & -1\end{pmatrix}\begin{pmatrix}-3\\6\\0\end{pmatrix} = 3/2,$$

and we get

	x_1	x_3	x_4	ξ_1	ξ_2	ξ_3
$1/2$	0	$3/2$	0	$-5/2$	$-3/2$	0
$x_4 = 1/2$	0	$-9/2$	1	$1/2$	$-1/2$	0
$x_1 = 1/2$	1	$1/2$	0	$1/6$	$1/6$	0
$\xi_3 = 1/2$	0	$3/2$	0	$-3/2$	$-1/2$	1

.

By analyzing the top row of reduced cost, we see that $j^+ = 2$. Furthermore, since $\min\{x_1/(1/2), \xi_3/(3/2)\} = \xi_3/(3/2) = 1/3$, we let $k^- = 6$, $K = (3, 1, 2)$, and pivot along the red circled 3/2 to obtain

	x_1	x_3	x_4	ξ_1	ξ_2	ξ_3
0	0	0	0	-1	-1	-1
$x_4 = 2$	0	0	1	-4	-2	3
$x_1 = 1/3$	1	0	0	$2/3$	$1/3$	$-1/3$
$x_3 = 1/3$	0	1	0	-1	$-1/3$	$2/3$

.

Since the upper left corner of the final tableau is zero and the reduced costs are all ≤ 0, we are finally finished. Then $y = (19/3\ 8/3\ -14/3)$ is an optimal solution of (D), but more importantly $(x_1, x_2, x_3, x_4) = (1/3, 0, 1/3, 2)$ is an optimal solution for our original linear program and provides an optimal value of $-10/3$.

The primal-dual algorithm for linear programming doesn't seem to be the favorite method to solve linear programs nowadays. But it is important because its basic principle, to use a restricted (simpler) primal problem involving an objective function with fixed weights, namely 1, and the dual problem to provide feedback to the primal by improving the objective function of the dual, has led to a whole class of combinatorial algorithms (often approximation algorithms) based on the primal-dual paradigm. The reader will get a taste of this kind of algorithm by consulting Papadimitriou and Steiglitz [47], where it is explained how classical algorithms such as Dijkstra's algorithm for the shortest path problem, and Ford and Fulkerson's algorithm for max flow can be derived from the primal-dual paradigm.

9.7 Summary

The main concepts and results of this chapter are listed below:

- Strictly separating hyperplane.
- Farkas–Minkowski proposition.
- Farkas lemma, version I, Farkas lemma, version II.
- Distance of a point to a subset.
- Dual linear program, primal linear program.
- Dual variables, primal variables.
- Complementary slackness conditions.
- Dual simplex algorithm.
- Primal-dual algorithm.
- Restricted primal linear program.

9.8 Problems

Problem 9.1. Let $(v_1, \ldots, v_n)$ be a sequence of n vectors in $\mathbb{R}^d$ and let V be the $d \times n$ matrix whose j-th column is v_j. Prove the equivalence of the following two statements:

(a) There is no nontrivial positive linear dependence among the v_j, which means that there is no nonzero vector, $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, with $y_j \geq 0$ for $j = 1, \ldots, n$, so that

$$y_1 v_1 + \cdots + y_n v_n = 0$$

or equivalently, $Vy = 0$.

(b) There is some vector, $c \in \mathbb{R}^d$, so that $c^\top V > 0$, which means that $c^\top v_j > 0$, for $j = 1, \ldots, n$.

Problem 9.2. Check that the dual in maximization form (D'') of the Dual Program (D') (which is the dual of (P) in maximization form),

$$\begin{aligned} &\text{maximize} && -b^\top y^\top \\ &\text{subject to} && -A^\top y^\top \leq -c^\top \text{ and } y^\top \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$, gives back the Primal Program (P).

Problem 9.3. In a General Linear Program (P) with n primal variables $x_1, \ldots, x_n$ and objective function $\sum_{j=1}^n c_j x_j$ (to be maximized), the m constraints are of the form

$$\begin{aligned} \sum_{j=1}^n a_{ij} x_j &\leq b_i, \\ \sum_{j=1}^n a_{ij} x_j &\geq b_i, \\ \sum_{j=1}^n a_{ij} x_j &= b_i, \end{aligned}$$

for $i = 1, \ldots, m$, and the variables x_j satisfy an inequality of the form

$$\begin{aligned} x_j &\geq 0, \\ x_j &\leq 0, \\ x_j &\in \mathbb{R}, \end{aligned}$$

for $j = 1, \ldots, n$. If $y_1, \ldots, y_m$ are the dual variables, show that the dual program of the linear program in standard form equivalent to (P) is equivalent to the linear program whose

objective function is $\sum_{i=1}^{m} y_i b_i$ (to be minimized) and whose constraints are determined as follows:

$$\text{if} \quad \left\{ \begin{array}{l} x_j \geq 0 \\ x_j \leq 0 \\ x_j \in \mathbb{R} \end{array} \right\}, \quad \text{then} \quad \left\{ \begin{array}{l} \displaystyle\sum_{i=1}^{m} a_{ij} y_i \geq c_j \\ \displaystyle\sum_{i=1}^{m} a_{ij} y_i \leq c_j \\ \displaystyle\sum_{i=1}^{m} a_{ij} y_i = c_j \end{array} \right\},$$

and

$$\text{if} \quad \left\{ \begin{array}{l} \displaystyle\sum_{j=1}^{n} a_{ij} x_j \leq b_i \\ \displaystyle\sum_{j=1}^{n} a_{ij} x_j \geq b_i \\ \displaystyle\sum_{j=1}^{n} a_{ij} x_j = b_i \end{array} \right\}, \quad \text{then} \quad \left\{ \begin{array}{l} y_i \geq 0 \\ y_i \leq 0 \\ y_i \in \mathbb{R} \end{array} \right\}.$$

Problem 9.4. Apply the procedure of Problem 9.3 to show that the dual of the (general) linear program

$$\begin{array}{ll} \text{maximize} & 3x_1 + 2x_2 + 5x_3 \\ \text{subject to} & \\ & 5x_1 + 3x_2 + x_3 = -8 \\ & 4x_1 + 2x_2 + 8x_3 \leq 23 \\ & 6x_1 + 7x_2 + 3x_3 \geq 1 \\ & x_1 \leq 4,\ x_3 \geq 0 \end{array}$$

is the (general) linear program:

$$\begin{array}{ll} \text{minimize} & -8y_1 + 23y_2 - y_3 + 4y_4 \\ \text{subject to} & \\ & 5y_1 + 4y_2 - 6y_3 + y_4 = 3 \\ & 3y_1 + 2y_2 - 7y_3 = 2 \\ & y_1 + 8y_2 - 3y_3 \geq 5 \\ & y_2,\ y_3,\ y_4 \geq 0. \end{array}$$

Problem 9.5. (1) Prove that the dual of the (general) linear program

$$\begin{array}{ll} \text{maximize} & cx \\ \text{subject to} & Ax = b \text{ and } x \in \mathbb{R}^n \end{array}$$

is

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA = c \text{ and } y \in \mathbb{R}^m. \end{aligned}$$

(2) Prove that the dual of the (general) linear program

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \geq b \text{ and } x \geq 0 \end{aligned}$$

is

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c \text{ and } y \leq 0. \end{aligned}$$

Problem 9.6. Use the complementary slackness conditions to confirm that

$$x_1 = 2,\ x_2 = 4,\ x_3 = 0,\ x_4 = 0,\ x_5 = 7,\ x_6 = 0$$

is an optimal solution of the following linear program (from Chavatal [18], Chapter 5):

$$\begin{aligned} &\text{maximize} \quad 18x_1 - 7x_2 + 12x_3 + 5x_4 + 8x_6 \\ &\text{subject to} \end{aligned}$$

$$\begin{aligned} 2x_1 - 6x_2 + 2x_3 + 7x_4 + 3x_5 + 8x_6 &\leq 1 \\ -3x_1 - x_2 + 4x_3 - 3x_4 + x_5 + 2x_6 &\leq -2 \\ 8x_1 - 3x_2 + 5x_3 - 2x_4 + 2x_6 &\leq 4 \\ 4x_1 + 8x_3 + 7x_4 - x_5 + 3x_6 &\leq 1 \\ 5x_1 + 2x_2 - 3x_3 + 6x_4 - 2x_5 - x_6 &\leq 5 \\ x_1, x_2, x_3, x_4, x_5, x_6 &\geq 0. \end{aligned}$$

Problem 9.7. Check carefully that the dual simplex method is equivalent to the simplex method applied to the dual program in maximization form.

Chapter 10

Basics of Combinatorial Topology

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, tetrahedra and more generally simplices, or even convex polytopes. We will begin by using simplices as building blocks.

The material presented in this chapter consists of the most basic notions of combinatorial topology, going back roughly to the 1900-1930 period and it is covered in nearly every algebraic topology book (certainly the "classics"). A classic text (slightly old fashion especially for the notation and terminology) is Alexandrov [1], Volume 1 and another more "modern" source is Munkres [45]. An excellent treatment from the point of view of computational geometry can be found is Boissonnat and Yvinec [12], especially Chapters 7 and 10. Another fascinating book covering a lot of the basics but devoted mostly to three-dimensional topology and geometry is Thurston [63].

One of the main goals of this chapter is to define a discrete (combinatorial) analog of the notion of a topological manifold (with or without boundary). The key for doing this is to define a combinatorial notion of nonsingularity of a face, and technically this is achieved by defining the notions of star and link of a face. There are actually two variants of the notion of star: closed stars and open stars. It turns out that the notion of nonsingularity is captured well by defining a face to be nonsingular if its link is homeomorphic to a sphere or to a closed ball. It is intuitively clear that if every face is nonsingular then the open star of every face is a "nice" open set, either an open ball or the intersection of an open ball with a half space.

However, proving this fact rigorously takes a surprising amount of work and requires the introduction of new concepts such as the suspension of a complex and the join of complexes. Once again, our geometric intuition in dimension greater than three is very unreliable, and we have to resort to algebraic arguments involving induction to be on solid grounds.

10.1 Simplicial Complexes

Recall that a simplex is just the convex hull of a finite number of affinely independent points. We also need to define faces, the boundary, and the interior of a simplex.

Definition 10.1. 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, $0 \leq i \leq n$; the simplex σ is often denoted by $(a_0, \ldots, a_n)$. We call n the *dimension* of the n-simplex σ, and the points $a_0, \ldots, a_n$ are the *vertices* of σ; we denote the set of vertices $\{a_0, \ldots, a_n\}$ by $\mathrm{vert}(\sigma)$. 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* $\mathrm{Int}\,\sigma = \sigma - \partial\sigma$ of σ. The interior $\mathrm{Int}\,\sigma$ of σ is sometimes called an *open simplex*. See Figure 10.1.

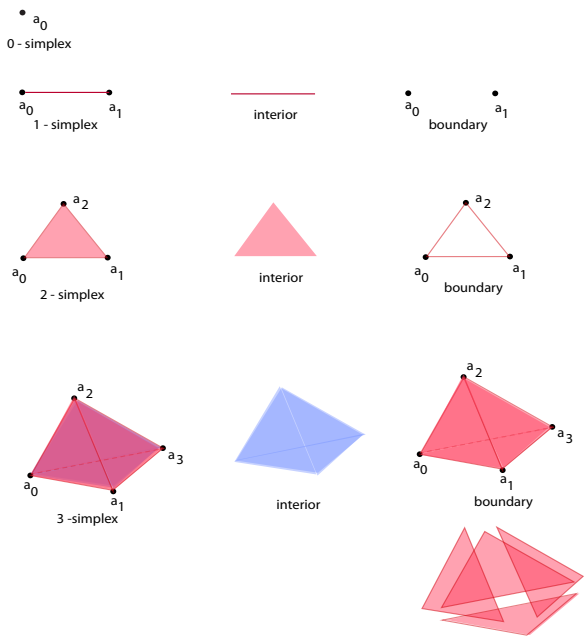


Figure 10.1: Illustrations of low-dimensional simplicies in $\mathbb{E}^3$, along with their corresponding interior and boundaries. The boundary of the 3-simplex (solid tetrahedron) consist of four triangles.

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 $\sigma \preceq \tau$.

We now state a number of properties of simplices, whose proofs are left as an exercise. 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 $\operatorname{Int}\sigma$ of a simplex is convex, arcwise connected, open, and σ is the closure of $\operatorname{Int}\sigma$.

We now put simplices together to form more complex shapes, following Munkres [45]. The intuition behind the next definition is that the building blocks should be "glued cleanly."

Definition 10.2. 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 10.2 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.

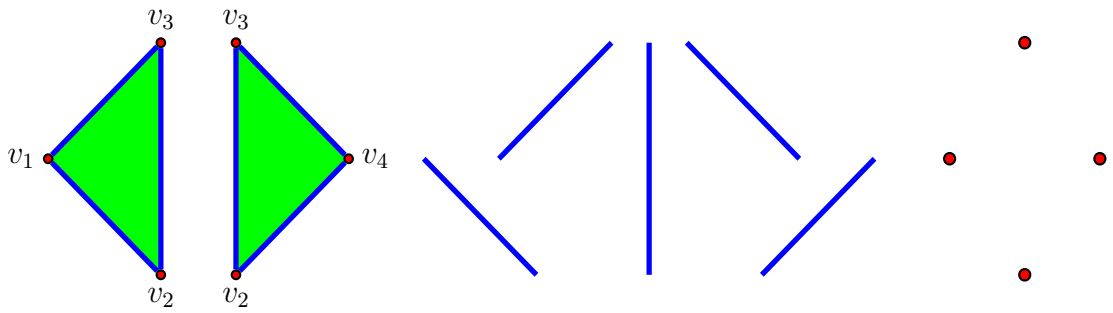


Figure 10.2: A set of simplices forming a complex

2. Unlike the situation for polyhedra, where all faces are external in the sense that they belong to the boundary of the polyhedron, the situation for simplicial complexes is more subtle; a face of a simplicial complex can be internal or external. For example, the 1-simplex (v_2, v_3) for the simplicial complex shown in Figure 10.2 is internal, but the the 1-simplex (v_1, v_2) is external. If we consider the simplicial complex consisting of the faces of a tetrahedron, then every edge (1-simplex) is internal. However, if we consider the simplicial complex consisting of a (solid) tetrahedron, then its facets (2-simplices) and edges (1-simplices) are external. These matters will be clarified in Definition 10.7.

3. 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.

4. 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 10.2 are shown in Figure 10.3. 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.

Some "legal" simplicial complexes are shown in Figure 10.5.

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.

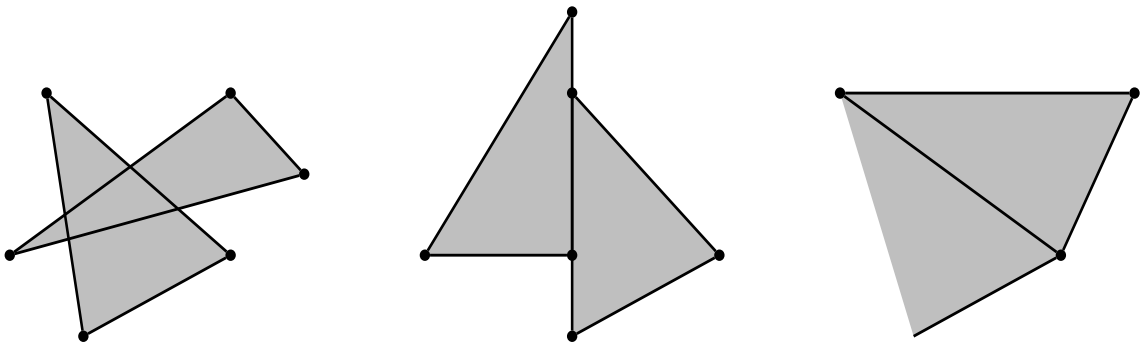


Figure 10.3: Collections of simplices not forming a complex

The resulting topological space $|K|$ is called the *geometric realization of* K. The geometric realization of the complex from Figure 10.2 is shown in Figure 10.4.

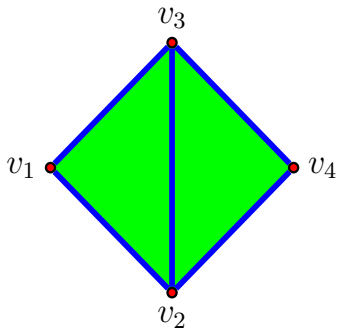


Figure 10.4: The geometric realization of the complex of Figure 10.2

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.

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*. Unfortunately the term "polytope" is overloaded since the polytopes induced by simplicial complexes are generally not convex. Consequently, if we use the term polytope for the objects defined in Chapter 5, we should really say "convex polytope" to avoid ambiguity. 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

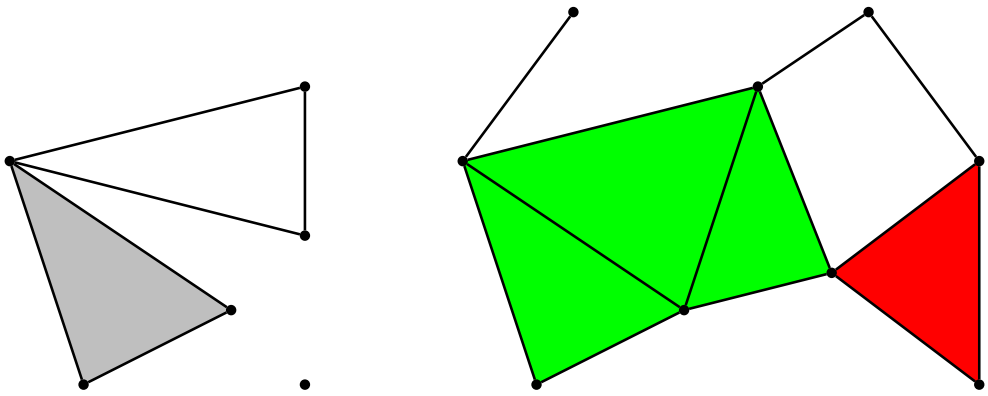


Figure 10.5: Examples of simplicial complexes

is identical to the topology induced from $\mathbb{E}^m$. Also, 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 10.3. 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. See Figure 10.6. 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). See Figure 10.7. A complex is *connected* iff $|K|$ is connected.

It 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 10.8, the complex on the left is not pure but the complex on the right is pure of dimension 2.

10.2 Nonsingular Faces; Stars and Links

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 "singu-

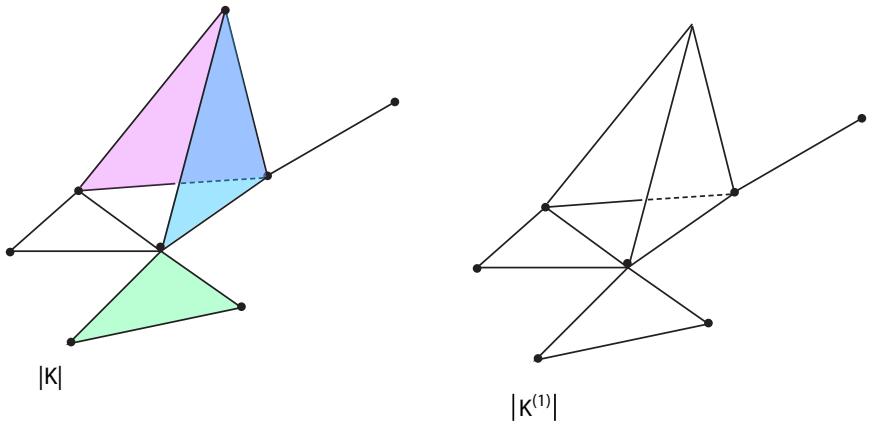


Figure 10.6: The one skeleton of a two-dimensional complex.

larities" such as the vertex v in Figure 10.8(b). The notions star and link of a face provide a technical way to deal with singularities.

Definition 10.4. 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, that is,

$$\mathrm{St}(\sigma) = \{s \in K \mid (\exists \tau \in K)(\sigma \preceq \tau \quad \text{and} \quad s \preceq \tau)\}.$$

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, that is,

$$\mathrm{Lk}(\sigma) = \{\tau \in K \mid \tau \in \mathrm{St}(\sigma) \quad \text{and} \quad \sigma \cap \tau = \emptyset\}.$$

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 10.9 shows

(a) A complex (on the left).

(b) The star of the vertex v, indicated in mint green and the link of v, shown as thicker red 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 [45], besides defining the complex $\mathrm{St}(\sigma)$, it is useful to introduce the open star of σ.

Definition 10.5. Given a complex K, for any simplex σ in K, the *open star* of σ, denoted $\mathrm{st}(\sigma)$, is 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.

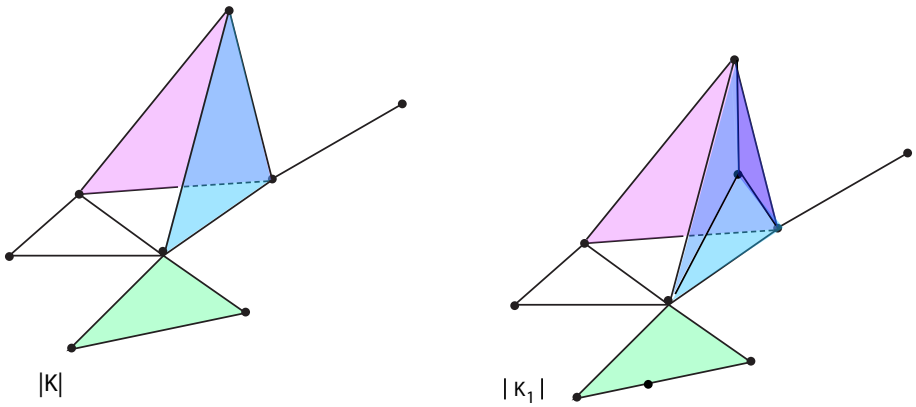


Figure 10.7: The complex K_1 is a subdivision of the two-dimensional complex K.

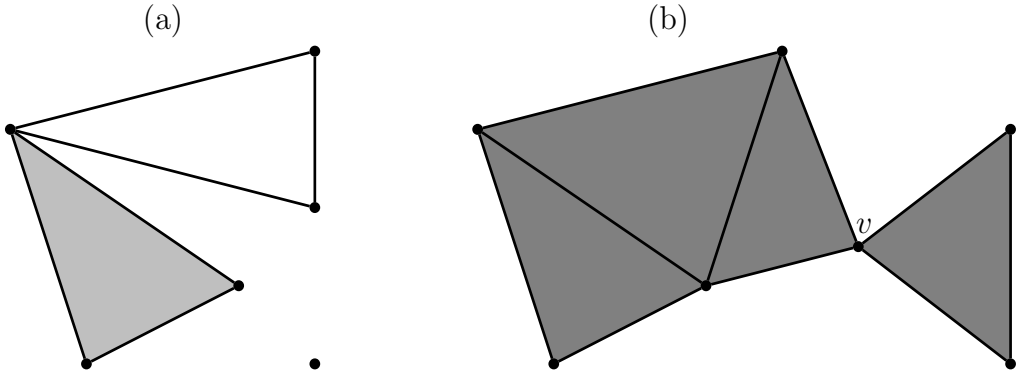


Figure 10.8: (a) A complex that is not pure. (b) A pure complex

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 $\overline{\mathrm{st}(\sigma)} - \mathrm{st}(\sigma)$! See Figures 10.10 and 10.11.

One of the nice properties of the open star $\mathrm{st}(\sigma)$ of σ is that it is open. To see this, observe that for any point $a \in |K|$, there is a unique smallest simplex $\sigma = (v_0, \ldots, v_k)$ such

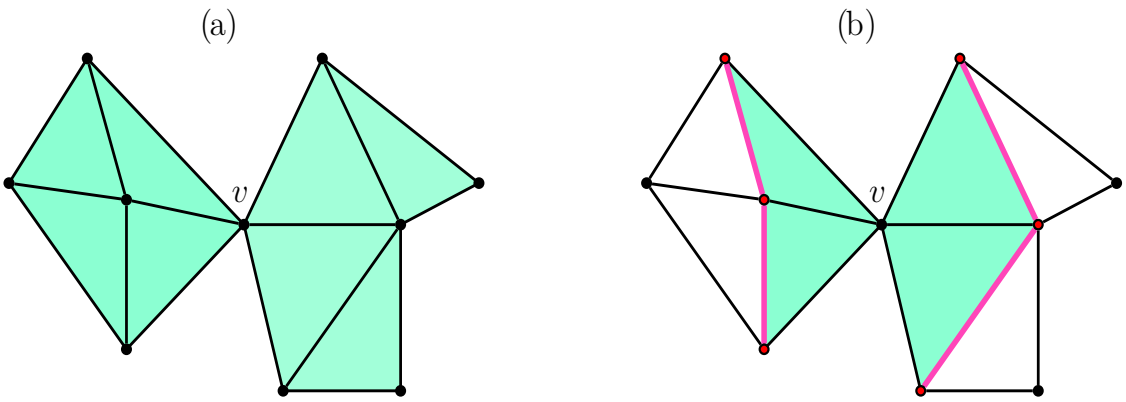


Figure 10.9: (a) A complex. (b) Star and Link of v.

that $a \in \mathrm{Int}(\sigma)$, that is, such that

$$a = \lambda_0 v_0 + \cdots + \lambda_k v_k$$

with $\lambda_i > 0$ for all i, with $0 \leq i \leq k$ (and of course, $\lambda_0 + \cdots + \lambda_k = 1$). (When $k = 0$, we have $v_0 = a$ and $\lambda_0 = 1$.) For every arbitrary vertex v of K, we define $t_v(a)$ by

$$t_v(a) = \begin{cases} \lambda_i & \text{if } v = v_i, \text{ with } 0 \leq i \leq k, \\ 0 & \text{if } v \notin \{v_0, \ldots, v_k\}. \end{cases}$$

Using the above notation, observe that

$$\mathrm{st}(v) = \{a \in |K| \mid t_v(a) > 0\},$$

and thus, $|K| - \mathrm{st}(v)$ is the union of all the faces of K that do not contain v as a vertex, obviously a closed set; see Figure 10.12. Thus, $\mathrm{st}(v)$ is open in $|K|$. It is also quite clear that $\mathrm{st}(v)$ is path connected. Moreover, for any k-face σ of K, if $\sigma = (v_0, \ldots, v_k)$, then

$$\mathrm{st}(\sigma) = \{a \in |K| \mid t_{v_i}(a) > 0, \quad 0 \leq i \leq k\},$$

that is,

$$\mathrm{st}(\sigma) = \mathrm{st}(v_0) \cap \cdots \cap \mathrm{st}(v_k).$$

Consequently, $\mathrm{st}(\sigma)$ is open and path connected, as illustrated in Figure 10.13.

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, (which is illustrated in Figure 10.14), consider the boundary of a tetrahedron and the star of a facet (a 2-simplex).

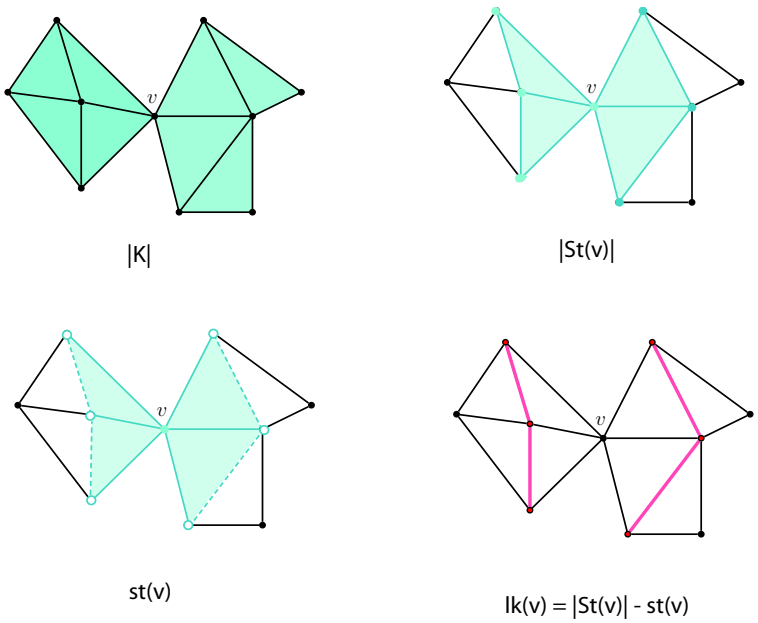


Figure 10.10: For the pure 2-dimensional complex $|K|$, an illustration of $|\mathrm{St}(v)|$, $\mathrm{st}(v)$, and $\mathrm{lk}(v)$.

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). The notion of link allows us to define precisely what we mean by a nonsingular face.

Definition 10.6. 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 $\mathrm{lk}(\sigma) \cong S^{d-k-1}$ or $\mathrm{lk}(\sigma) \cong \overline{B}^{d-k-1}$, where $\cong$ means homeomorphic.

In Figure 10.9, note that the link of v is not homeomorphic to S^1 or $\overline{B}^1$, so v is singular.

Given a pure complex, it is necessary to distinguish between two kinds of faces.

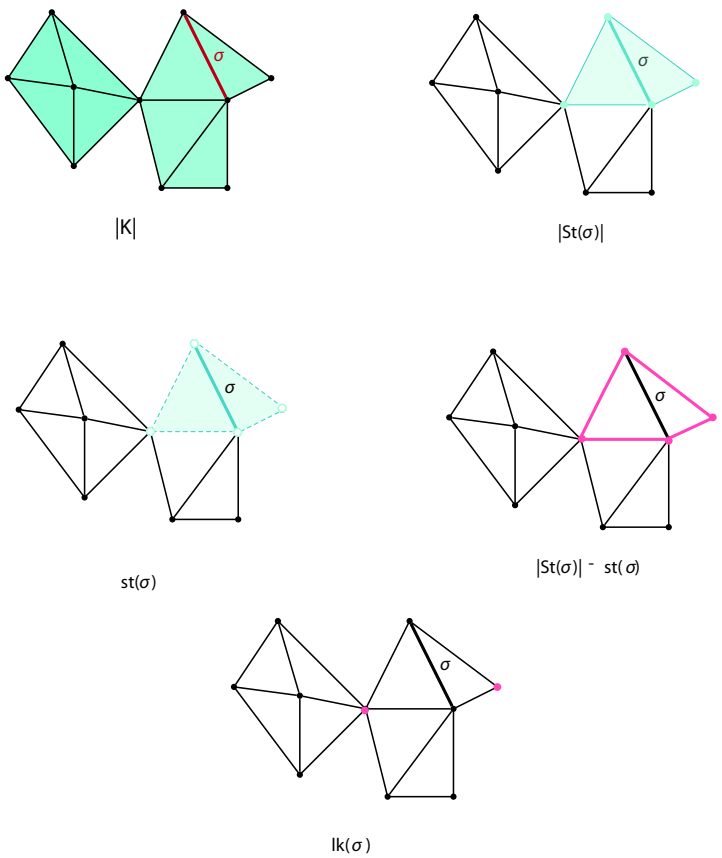


Figure 10.11: Given the edge σ in the pure 2-dimensional complex $|K|$, observe that $\mathrm{lk}(\sigma) \subset \overline{\mathrm{st}(\sigma)} - \mathrm{st}(\sigma)$.

Definition 10.7. 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 10.1. *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* $\mathrm{bd}(K)$*.*

Proof. Let F be any facet of K containing σ. We may assume that $F = (v_0, \ldots, v_{d-1})$ and

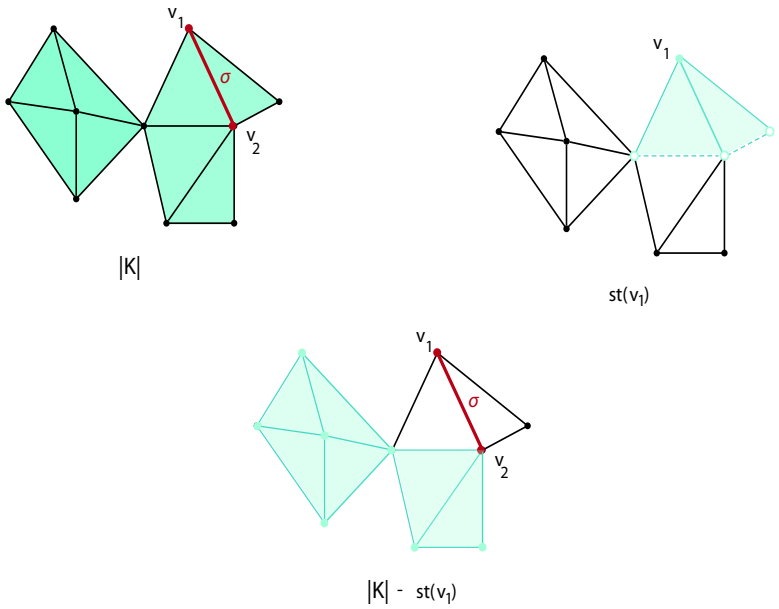


Figure 10.12: The construction of the closed set $|K| - \mathrm{st}(v_1)$.

$\sigma = (v_0, \ldots, v_k)$, in which case, $F' = (v_{k+1}, \ldots, v_{d-1})$ is a $(d-k-2)$-face of K and by definition of $\mathrm{Lk}(\sigma)$, we have $F' \in \mathrm{Lk}(\sigma)$. Now, every cell (i.e., d-simplex) s containing F is of the form $s = \mathrm{conv}(F \cup \{v\})$ for some vertex v, and $s' = \mathrm{conv}(F' \cup \{v\})$ is a $(d-k-1)$-face in $\mathrm{Lk}(\sigma)$ containing F'. Consequently, F' is an external face of $\mathrm{Lk}(\sigma)$ iff F is an external facet of K, establishing the proposition. The second statement follows immediately from the proof of the first. □

Proposition 10.1 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.

The main goal of the rest of this section is to show that if K is a pure complex of dimension d and if all its k-faces are nonsingular ($0 \leq k \leq d-1$), then $\mathrm{lk}(\sigma)$ is either homeomorphic to B^d or to $\overline{B}^d - \overline{B}^{d-1}$. As a consequence, the geometric realization $|K|$ of K is a manifold.

Although the above facts are easy to check for $d = 1, 2$, and in some simple cases for $d = 3$, a rigorous proof requires a fair amount of work and the introduction of several new concepts. The key point is that we need to express $\mathrm{St}(\sigma)$ in terms of $\mathrm{Lk}(\sigma)$, and for this, we need the notion of join of complexes, two special cases of which are the notion of cone and of suspension. These two notions allow building the sphere S^{d+1} and the closed ball $\overline{B}^{d+1}$ from the sphere S^d and the closed ball $\overline{B}^d$, and allow an inductive argument on d. There are many technical details which we will omit to simplify the exposition. Complete details and proofs can be found in Munkres [45] (Chapter 8, Section 62).

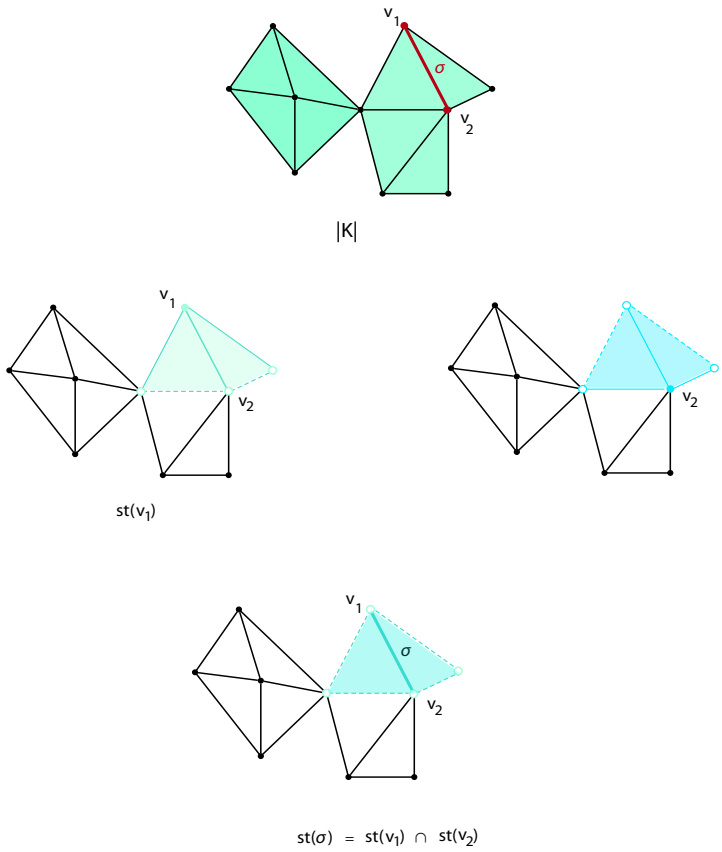


Figure 10.13: Given the edge σ in the pure 2-dimensional complex $|K|$, $\mathrm{st}(\sigma) = \mathrm{st}(v_1) \cap \mathrm{st}(v_2)$.

We begin with yet another notion of cone.

Definition 10.8. 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 $v * K = v$. See Figure 10.15

It is not hard to check that $v * K$ is indeed a complex of dimension $d + 1$ containing K as a subcomplex.

Remark: Unfortunately, the word "cone" is overloaded. It might have been better to use the locution *pyramid* instead of cone as some authors do (for example, Ziegler). However, since we have been following Munkres [45], a standard reference in algebraic topology, we decided to stick with the terminology used in that book, namely, "cone."

If σ is a simplex in a complex K, we will need to express $\mathrm{St}(\sigma)$ in terms of σ and its link $\mathrm{Lk}(\sigma)$, and $|\mathrm{St}(\sigma)| - \mathrm{st}(\sigma)$ in terms of $\partial\sigma$ and $\mathrm{Lk}(\sigma)$. For this, we will need a generalization

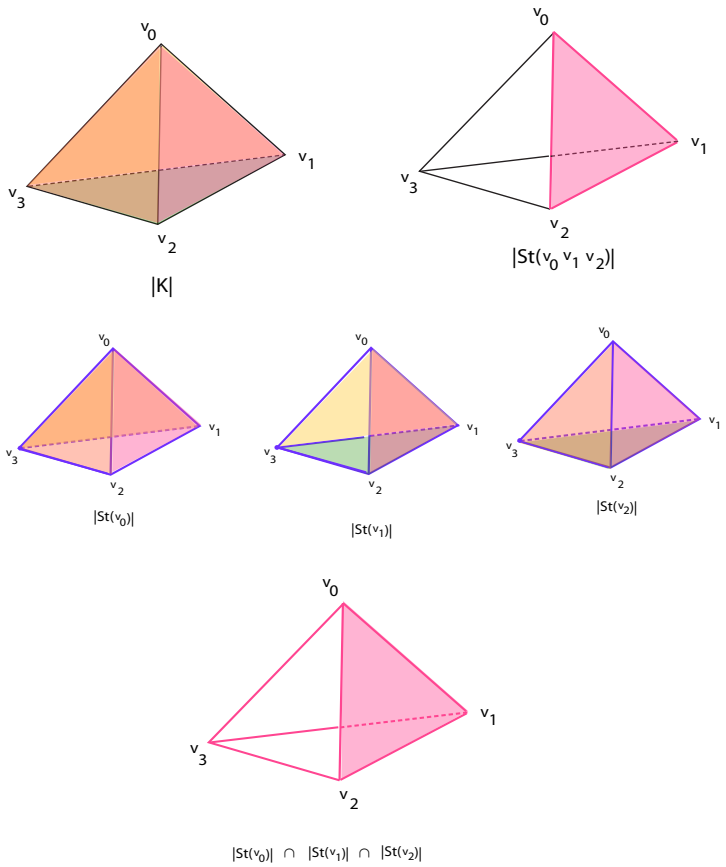


Figure 10.14: Let $|K|$ be the boundary of the solid tetrahedron. The star of a triangular face is itself and contains only three edges. It is not the intersection of the star of its vertices, since the star of a vertex contains all six edges of the tetrahedron.

of the above notion of cone to two simplicial complexes K and L, called the join of two complexes.

Definition 10.9. Given any two disjoint nonempty complexes K and L in $\mathbb{E}^n$ such that $\dim(K) + \dim(L) \leq n - 1$, if for any simplex $\sigma = (v_0, \ldots, v_h)$ in K and any simplex $\tau = (w_0, \ldots, w_k)$ in L, the points $(v_0, \ldots, v_h, w_0, \ldots, w_k)$ are affinely independent, then we define $\sigma * \tau$ as the simplex

$$\sigma * \tau = (v_0, \ldots, v_h, w_0, \ldots, w_k);$$

more rigorously, $\sigma * \tau$ is the $(h+k+1)$-simplex spanned by the points $(v_0, \ldots, v_h, w_0, \ldots, w_k)$. If the collection of all the simplicies $\sigma * \tau$ and their faces is a simplicial complex, then this complex is denoted by $K * L$ and is called the *join* of K and L.

Note that if $K * L$ is a complex, then its dimension is $\dim(K) + \dim(L) + 1$, which implies that $\dim(K) + \dim(L) \leq n - 1$.

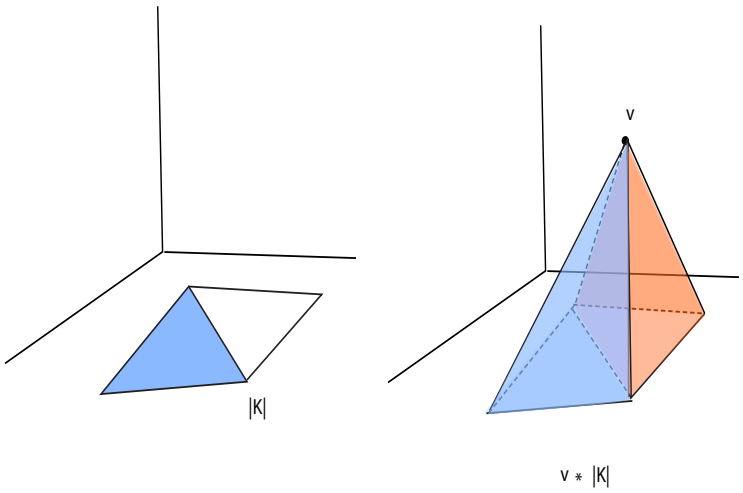


Figure 10.15: On the left is the two-dimensional planar complex $|K|$. On the right is the geometric realization of $|v*K|$. It consists of a solid blue tetrahedron and a peach tetrahedral shell.

Observe that a cone $v*L$ corresponds to the special case where K is a complex consisting of the single vertex v. If $K = \{v_0, v_1\}$ is the complex consisting of two distinct vertices (with no edge between them), and if $\{v_0, v_1\} * L$ is a complex, then it is called a *suspension* of L and it is denoted by $S(L)$ or $\mathrm{susp}(L)$. The suspension of L is the complex consisting of the union of the two cones $v_0 * L$ and $v_1 * L$.

Two problems immediately come to mind:

(1) Characterize the geometric realization $|K * L|$ of the join $K * L$ of two complexes K and L in terms of the geometric realizations $|K|$ and $|L|$ of K and L, if $K * L$ is indeed a complex.

(2) Find a sufficient condition of $|K|$ and $|L|$ that implies that $K * L$ is a complex.

The following proposition gives answers to these problems and gives a necessary and sufficient condition for $K * L$ to be a complex. The proof is quite technical and not very illuminating so we refer the reader to Munkres [45] (Chapter 8, Lemma 62.1).

Proposition 10.2. *Let K and L be any two disjoint nonempty complexes in $\mathbb{E}^n$ such that $\dim(K) + \dim(L) \leq n - 1$.*

(a) If $K * L$ is a complex, then its geometric realization $|K * L|$ is the union of all the closed line segments $[x, y]$ joining some point x in $|K|$ to some point y in $|L|$. Two such line segments intersect in a most a common endpoint.

(b) Conversely, if every pair of line segments joining points of $|K|$ and points of $|L|$ intersect in at most a common endpoint, then $K * L$ is a complex.

Proposition 10.2 shows that $|K * L|$ can be expressed in terms of the realizations of cones of the form $v * L$, where $v \in K$, as

$$|K * L| = \bigcup_{v \in |K|} |v * L|.$$

A few more technical propositions, all proved in Munkres [45] (see Chapter 8, Section 62), will be needed.

A surjective function $f\colon X \to Y$ between two topological spaces X and Y is called a *quotient map* if a subset V of Y is open iff $f^{-1}(V)$ is open in X. A quotient map is automatically continuous.

Proposition 10.3. *Suppose $K * L$ is a well-defined complex where K and L are finite complexes (it suffices to assume that K is locally finite). Then the map*

$$\pi\colon |K| \times |L| \times [0,1] \to |K * L|$$

given by

$$\pi(x, y, t) = (1 - t)x + ty$$

is a quotient map. For every $x \in |K|$ and every $y \in |L|$, the map π collapes $\{x\} \times |L| \times \{0\}$ to the point x and $|K| \times \{y\} \times \{1\}$ to the point y. Otherwise, π is injective.

Using Proposition 10.3 we can prove the following "obvious" proposition which turns out to be very handy.

Proposition 10.4. *Suppose $K * L$ and $M * N$ are well-defined finite complexes (it suffices to assume that K is locally finite). If $|K| \cong |M|$ and $|L| \cong |N|$, then $|K * L| \cong |M * N|$.*

The next proposition follows immediately from the definitions.

Proposition 10.5. *Let J, K, L be complexes in $\mathbb{E}^n$. If $J * K$ and $(J * K) * L$ are well-defined, then $K * J$ and $J * (K * L)$ are also well-defined and $J * K = K * J$, $(J * K) * L = J * (K * L)$.*

The following proposition shown in Munkres [45] (Chapter 8, Lemma 62.6) is crucial. The proof actually follows pretty much from the definitions.

Proposition 10.6. *For any complex K of dimension d and any k-simplex $\sigma \in K$ ($0 \leq k \leq d - 1$), we have*

$$\mathrm{St}(\sigma) = \sigma * \mathrm{Lk}(\sigma),$$

and

$$\mathrm{st}(\sigma) \cong |\mathrm{St}(\sigma)| - |\partial\sigma * \mathrm{Lk}(\sigma)|.$$

By convention, $\sigma * \emptyset = \sigma$ if $\mathrm{Lk}(\sigma) = \emptyset$, and $\emptyset * \mathrm{Lk}(\sigma) = \mathrm{Lk}(\sigma)$ if $\partial\sigma = \emptyset$. See Figures 10.16 and 10.17.

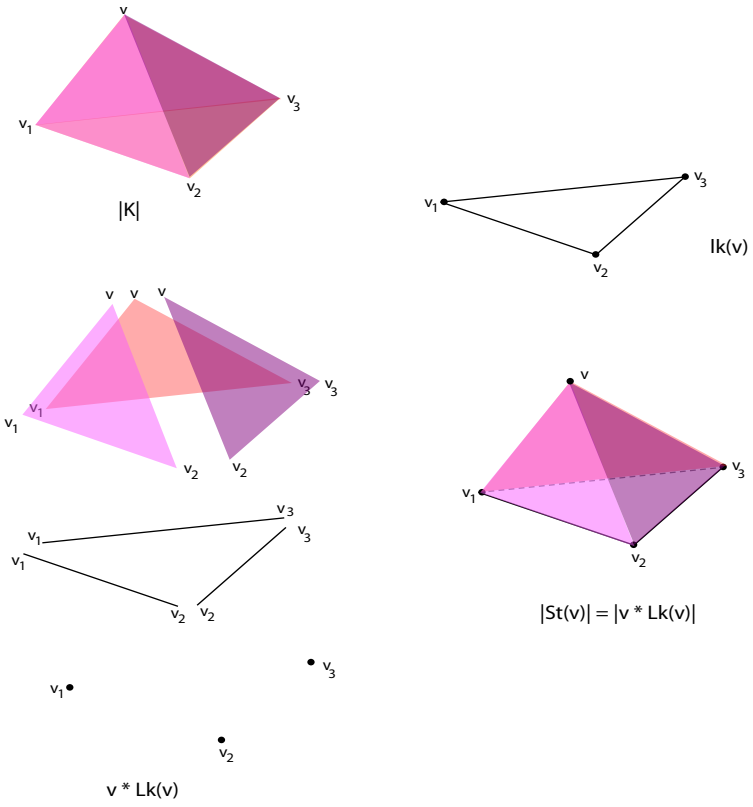


Figure 10.16: An illustration of the formula $\mathrm{St}(v) \cong v * \mathrm{Lk}(v)$ when $|K|$ is the two-dimensional tetrahedral shell.

Figure 10.18 shows a 3-dimensional complex. The link of the edge (v_6, v_7) is the pentagon $P = (v_1, v_2, v_3, v_4, v_5) \cong S^1$. The link of the vertex v_7 is the cone $v_6 * P \cong \overline{B}^2$. The link of (v_1, v_2) is $(v_6, v_7) \cong \overline{B}^1$ 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 $\overline{B}^2$.

The following technical propositions are needed to show that if K is any pure complex of dimension d, nonsingularity of all the faces implies that every open star is an open subset homeomorphic either to B^d or to $B^d \cap \mathbb{H}^d$, where

$$\mathbb{H}^d = \{(x_1, \ldots, x_d) \in \mathbb{R}^d \mid x_d \geq 0\}.$$

The *standard simplex* Δ^d is the convex subset of $\mathbb{R}^{d+1}$ given by

$$\Delta^d = \{(t_0, \ldots, t_d) \in \mathbb{R}^{d+1} \mid t_0 + \cdots + t_d = 1,\ t_i \geq 0\}.$$

It is easy to show that

$$|\Delta^d| \cong \overline{B}^d, \quad |\partial\Delta^{d+1}| \cong S^d, \quad \text{for all } d \geq 0;$$

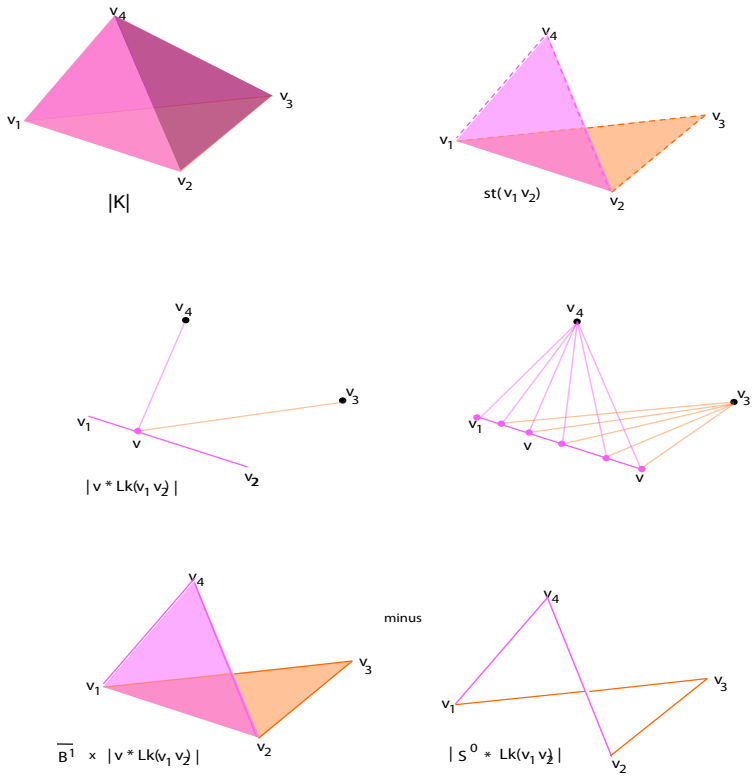


Figure 10.17: An illustration of the formula $\mathrm{st}(\sigma) \cong |\sigma * \mathrm{Lk}(\sigma)| - |S^0 * \mathrm{Lk}(\sigma)|$ when $|K|$ is the two-dimensional tetrahedral shell and $\sigma = (v_1, v_2)$.

see Munkres [45] (Chapter 1, Lemma 1.1). The following formulae are also easy to show but they are essential to carry out induction on the dimension of spheres or (closed) balls.

Proposition 10.7. *The following homeomorphisms hold for all $d \geq 0$:*

$$\begin{aligned} |v * \partial\Delta^{d+1}| &\cong |\partial\Delta^{d+1} * v| \cong \overline{B}^{d+1} \\ |v * \Delta^{d}| &\cong |\Delta^{d} * v| \cong \overline{B}^{d+1} \\ |\partial\Delta^{d+1} * \{v_0, v_1\}| &\cong S^{d+1} \\ |\Delta^{d} * \{v_0, v_1\}| &\cong \overline{B}^{d+1}, \end{aligned}$$

for any points $v, v_0 \neq v_1$ not in Δ^d.

Informally, the first formula says that a cone over the sphere S^d is homeomorphic to the closed ball $\overline{B}^{d+1}$, the second formula says that a cone over the closed ball $\overline{B}^{d}$, is homeomorphic to the closed ball $\overline{B}^{d+1}$, the third formula says that the suspension of the sphere S^d is

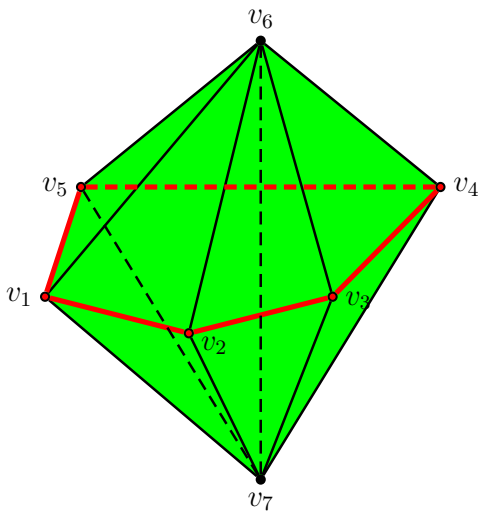


Figure 10.18: More examples of links and stars

homeomorphic to the sphere S^{d+1}, and the fourth formula says that the suspension of the closed ball $\overline{B}^d$ is homeomorphic to the closed ball $\overline{B}^{d+1}$.

Proposition 10.8. *For every $d \geq 1$ and every k-simplex σ ($0 \leq k \leq d-1$), we have*

$$|\sigma * \partial\Delta^{d-k}| \cong \overline{B}^d$$
$$|\partial\sigma * \partial\Delta^{d-k}| \cong S^{d-1}.$$

Proof. We proceed by induction on $d \geq 1$. For the base case $d = 1$, we must have $k = 0$ so $\sigma = v$ and $\partial\sigma = \emptyset$ for some vertex v, and then by Proposition 10.7

$$|v * \partial\Delta^1| \cong \overline{B}^1,$$

and

$$|\emptyset * \partial\Delta^1| = |\partial\Delta^1| \cong S^0.$$

For the induction step for the first formula, we use Proposition 10.7 which says that

$$|\partial\Delta^{d-k} * \{v_0, v_1\}| \cong S^{d-k+1} \cong |\partial\Delta^{d-k+1}|.$$

For any k-simplex σ with $0 \leq k \leq d-1$, by Proposition 10.4 we have

$$\begin{aligned} |\sigma * \partial\Delta^{d-k+1}| &\cong |\sigma * (\partial\Delta^{d-k} * (\{v_0, v_1\})| \\ &\cong |(\sigma * \partial\Delta^{d-k}) * \{v_0, v_1\}|. \end{aligned}$$

By the induction hypothesis, we have

$$|\sigma * \partial\Delta^{d-k}| \cong \overline{B}^d = |\Delta^d|,$$

so by Proposition 10.4 and Proposition 10.7, we have

$$\begin{aligned}|\sigma * \partial\Delta^{d-k+1}| &\cong |(\sigma * \partial\Delta^{d-k}) * \{v_0, v_1\}| \\ &\cong |\Delta^d * \{v_0, v_1\}| \\ &\cong \overline{B}^{d+1}.\end{aligned}$$

For a d-simplex σ, since $|\sigma| \cong |\Delta^d|$ and $|\partial\Delta^1| = |\{v_0, v_1\}|$, by Proposition 10.4 and Proposition 10.7, we have

$$\begin{aligned}|\sigma * \partial\Delta^1| &\cong |\Delta^d * \{v_0, v_1\}| \\ &\cong \overline{B}^{d+1}.\end{aligned}$$

This concludes the induction step for the first formula and proves that

$$|\sigma * \partial\Delta^{d-k}| \cong \overline{B}^d.$$

For the second formula, if $k = 0$ then $\sigma = v$ is a vertex and $\partial\sigma = \emptyset$ so

$$|\emptyset * \partial\Delta^d| = |\partial\Delta^d| \cong S^{d-1}.$$

For any k-simplex σ with $1 \leq k \leq d-1$, by Proposition 10.4 we have

$$\begin{aligned}|\partial\sigma * \partial\Delta^{d-k+1}| &\cong |\partial\sigma * (\partial\Delta^{d-k} * (\{v_0, v_1\})| \\ &\cong |(\partial\sigma * \partial\Delta^{d-k}) * \{v_0, v_1\}|.\end{aligned}$$

By the induction hypothesis, we have

$$|\partial\sigma * \partial\Delta^{d-k}| \cong S^{d-1} = |\partial\Delta^d|,$$

so by Proposition 10.4 and Proposition 10.7, we have

$$\begin{aligned}|\partial\sigma * \partial\Delta^{d-k+1}| &\cong |(\partial\sigma * \partial\Delta^{d-k}) * \{v_0, v_1\}| \\ &\cong |\partial\Delta^d * \{v_0, v_1\}| \\ &\cong S^d.\end{aligned}$$

For a d-simplex σ, since $|\partial\sigma| \cong |\partial\Delta^d|$ and $|\partial\Delta^1| \cong |\{v_0, v_1\}|$, by Proposition 10.4 and Proposition 10.7, we have

$$\begin{aligned}|\partial\sigma * \partial\Delta^1| &\cong |\partial\sigma * \{v_0, v_1\}| \\ &\cong |\partial\Delta^d * \{v_0, v_1\}| \\ &\cong S^d.\end{aligned}$$

This concludes the induction step for the second formula and proves that

$$|\partial\sigma * \partial\Delta^{d-k}| \cong S^{d-1},$$

as claimed. □

Since $|\partial\Delta^{d-k}| \cong S^{d-k-1}$, with a slight abuse of notation the formulae of Proposition 10.8 can be written as

$$|\sigma * S^{d-k-1}| \cong \overline{B}^d$$
$$|\partial\sigma * S^{d-k-1}| \cong S^{d-1}.$$

The following proposition is the counterpart of Proposition 10.8 for balls instead of spheres.

Proposition 10.9. *For every $d \geq 1$ and every k-simplex σ ($0 \leq k \leq d-1$), we have*

$$|\sigma * \Delta^{d-k-1}| \cong \overline{B}^d$$
$$|\partial\sigma * \Delta^{d-k-1}| \cong \overline{B}^{d-1}.$$

Proof. We proceed by induction on $d \geq 1$. For the base case $d = 1$, we must have $k = 0$ so $\sigma = v$ for some vertex v, and then by Proposition 10.7

$$|v * \Delta^0| \cong \overline{B}^1,$$

and by definition

$$|\emptyset * \Delta^0| = |\Delta^0| \cong \overline{B}^0.$$

For the induction step for the first formula, we use Proposition 10.7 which says that

$$|\Delta^{d-k-1} * v| \cong \overline{B}^{d-k} = |\Delta^{d-k}|.$$

For any k-simplex σ with $0 \leq k \leq d-1$, by Proposition 10.4 we have

$$\begin{aligned} |\sigma * \Delta^{d-k}| &\cong |\sigma * (\Delta^{d-k-1} * v)| \\ &\cong |(\sigma * \Delta^{d-k-1}) * v|. \end{aligned}$$

By the induction hypothesis, we have

$$|\sigma * \Delta^{d-k-1}| \cong \overline{B}^d = |\Delta^d|,$$

so by Proposition 10.4 and Proposition 10.7, we have

$$\begin{aligned} |\sigma * \Delta^{d-k}| &\cong |(\sigma * \Delta^{d-k-1}) * v| \\ &\cong |\Delta^d * v| \\ &\cong \overline{B}^{d+1}. \end{aligned}$$

For a d-simplex σ, since $|\sigma| \cong |\Delta^d|$ and $|\Delta^0| = |v|$, by Proposition 10.4 and Proposition 10.7, we have

$$\begin{aligned} |\sigma * \Delta^0| &\cong |\Delta^d * v| \\ &\cong \overline{B}^{d+1}. \end{aligned}$$

This concludes the induction step for the first formula and proves that

$$|\sigma * \Delta^{d-k-1}| \cong \overline{B}^d.$$

For the second formula, if $k = 0$ then $\sigma = v$ is a vertex and $\partial\sigma = \emptyset$ so

$$|\emptyset * \Delta^{d-1}| = |\Delta^{d-1}| \cong \overline{B}^{d-1}.$$

For any k-simplex σ with $1 \leq k \leq d-1$, by Proposition 10.4 we have

$$\begin{aligned} |\partial\sigma * \Delta^{d-k}| &\cong |\partial\sigma * (\Delta^{d-k-1} * v)| \\ &\cong |(\partial\sigma * \Delta^{d-k-1}) * v|. \end{aligned}$$

By the induction hypothesis, we have

$$|\partial\sigma * \Delta^{d-k-1}| \cong \overline{B}^{d-1} = |\Delta^{d-1}|,$$

so by Proposition 10.4 and Proposition 10.7, we have

$$\begin{aligned} |\partial\sigma * \Delta^{d-k}| &\cong |(\partial\sigma * \Delta^{d-k-1}) * v| \\ &\cong |\Delta^{d-1} * v| \\ &\cong \overline{B}^d. \end{aligned}$$

For a d-simplex σ, since $|\partial\sigma| \cong |\partial\Delta^d|$ and $|\Delta^0| \cong |v|$, by Proposition 10.4 and Proposition 10.7, we have

$$\begin{aligned} |\partial\sigma * \Delta^0| &\cong |\partial\sigma * v| \\ &\cong |\partial\Delta^d * v| \\ &\cong \overline{B}^d. \end{aligned}$$

This concludes the induction step for the second formula and proves that

$$|\partial\sigma * \Delta^{d-k-1}| \cong \overline{B}^{d-1},$$

as claimed. □

Since $|\Delta^{d-k-1}| \cong \overline{B}^{d-k-1}$, with a slight abuse of notation the formulae of Proposition 10.9 can be written as

$$\begin{aligned} |\sigma * \overline{B}^{d-k-1}| &\cong \overline{B}^d \\ |\partial\sigma * \overline{B}^{d-k-1}| &\cong \overline{B}^{d-1}. \end{aligned}$$

Finally, we can prove that for any pure complex K of dimension d, nonsingularity of all the faces implies that the open star of any internal face is homeomorphic to B^d, and that the open star of any boundary face is homeomorphic to $\overline{B}^d - \overline{B}^{d-1}$. This result for pure complexes K without boundaries is stated in Thurston [63] (Chapter 3, Proposition 3.2.5).

Theorem 10.10. *Let K be any pure complex of dimension d. If every face of K is nonsingular, then $\mathrm{st}(\sigma) \cong B^d$ for every every internal k-face σ of K, and $\mathrm{st}(\sigma) \cong \overline{B}^d - \overline{B}^{d-1}$ for every every boundary k-face σ of K ($0 \leq k \leq d-1$).*

Proof. By Proposition 10.6, for any complex K of dimension d and any k-simplex $\sigma \in K$ ($0 \leq k \leq d-1$), we have

$$\mathrm{St}(\sigma) = \sigma * \mathrm{Lk}(\sigma),$$

and

$$\mathrm{st}(\sigma) \cong |\mathrm{St}(\sigma)| - |\partial\sigma * \mathrm{Lk}(\sigma)|.$$

If σ is an internal face then

$$|\mathrm{Lk}(\sigma)| \cong S^{d-k-1} \cong |\partial\Delta^{d-k}|,$$

so by Proposition 10.8 and Proposition 10.4

$$\begin{aligned}\mathrm{st}(\sigma) &\cong |\mathrm{St}(\sigma)| - |\partial\sigma * \mathrm{Lk}(\sigma)| \\ &\cong |\sigma * \partial\Delta^{d-k}| - |\partial\sigma * \partial\Delta^{d-k}| \\ &\cong \overline{B}^d - S^{d-1} \\ &\cong B^d.\end{aligned}$$

If σ is a boundary face then

$$|\mathrm{Lk}(\sigma)| \cong \overline{B}^{d-k-1} \cong |\Delta^{d-k-1}|,$$

so by Proposition 10.9 and Proposition 10.4

$$\begin{aligned}\mathrm{st}(\sigma) &\cong |\mathrm{St}(\sigma)| - |\partial\sigma * \mathrm{Lk}(\sigma)| \\ &\cong |\sigma * \Delta^{d-k-1}| - |\partial\sigma * \Delta^{d-k-1}| \\ &\cong \overline{B}^d - \overline{B}^{d-1},\end{aligned}$$

as claimed. □

Theorem 10.10 has the following corollary which shows that any pure complex for which every face is nonsingular is a manifold.

Proposition 10.11. *Let K be any pure complex of dimension d. If every face of K is nonsingular, then for every point $a \in |K|$, there is an open subset $U \subseteq |K|$ containing a such that if a does not belong to the boundary of $|K|$ then $U \cong B^d$, and if a belongs to the boundary of $|K|$ then $U \cong B^d \cap \mathbb{H}^d$.*

Proof. Any point $a \in |K|$ belongs to some simplex σ, so we proceed by induction on the dimension of σ. If σ is a vertex v, then by Proposition 10.10 we have $\mathrm{st}(v) \cong B^d$ or $\mathrm{st}(\sigma) \cong \overline{B}^d - \overline{B}^{d-1} \cong B^d \cap \mathbb{H}^d$. If σ is a simplex of dimension $k+1$, then any point $a \in \partial\sigma$ on the boundary of σ belongs to a simplex of dimension at most k, and the induction hypothesis implies that there is an open subset $U \subseteq |K|$ containing a such that $U \cong B^d$ or $U \cong B^d \cap \mathbb{H}^d$. Otherwise, a belongs to the interior of σ, and we conclude by Proposition 10.10 since $\mathrm{st}(\sigma) \cong B^d$ or $\mathrm{st}(\sigma) \cong \overline{B}^d - \overline{B}^{d-1} \cong B^d \cap \mathbb{H}^d$. □

Remark: Thurston states that Proposition 10.11 holds for pure complexes without boundaries under the weaker assumption that $\mathrm{lk}(v) \cong S^{d-1}$ for every vertex; see Thurston [63], Chapter 3, Proposition 3.2.5. A proof of the more general fact that if $\mathrm{lk}(v) \cong S^{d-1}$ or $\mathrm{lk}(v) \cong \overline{B}^{d-1}$ for every vertex then *every* face is nonsingular can be found in Stallings [56]; see Section 4.4, Proposition 4.4.12. The proof requires several technical lemmas and is quite involved.

Here are more useful propositions about pure complexes without singularities.

Proposition 10.12. *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).*

Proof. If $|K| \subseteq \mathbb{E}^d$, then this is an immediate consequence of the definition of a complex. Otherwise, consider $\mathrm{lk}(\sigma)$. By hypothesis, either $\mathrm{lk}(\sigma) \cong B^0$ or $\mathrm{lk}(\sigma) \cong S^0$. As $B^0 = \{0\}$, $S^0 = \{-1, 1\}$ and $\dim \mathrm{Lk}(\sigma) = 0$, we deduce that $\mathrm{Lk}(\sigma)$ has either one or two points, which proves that σ belongs to at most two d-simplices. □

Proposition 10.13. *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 $i = 0, \ldots, p-1$.*

Proof. We proceed by induction on d, using the fact that the links are connected for $d \geq 2$. □

Proposition 10.14. *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.*

Proof. Left as an exercise. □

10.3 Polyhedral Complexes

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 10.10. 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*.

Remark: Since the building blocks of a polyhedral complex are convex *polytopes* it might be more appropriate to use the term "polytopal complex" rather than "polyhedral complex" and some authors do that. On the other hand, most of the traditional litterature uses the terminology *polyhedral complex* so we will stick to it. There is a notion of complex where the building blocks are cones but these are called *fans*.

Every convex 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 5.7, every polytope σ is the convex hull of its vertices.

Let $\mathrm{vert}(\sigma)$ denote the set of vertices of σ. Then, we have the following crucial observation: Given any polyhedral complex K, for every point $x \in |K|$, there is a *unique* polytope $\sigma_x \in K$ such that $x \in \mathrm{Int}(\sigma_x) = \sigma_x - \partial\,\sigma_x$. We define a function $t\colon V \to \mathbb{R}_+$ that tests whether x belongs to the interior of any face (polytope) of K having v as a vertex as follows: For every vertex v of K,

$$t_v(x) = \begin{cases} 1 & \text{if } v \in \mathrm{vert}(\sigma_x) \\ 0 & \text{if } v \notin \mathrm{vert}(\sigma_x), \end{cases}$$

where σ_x is the unique face of K such that $x \in \mathrm{Int}(\sigma_x)$.

Now, just as in the simplicial case, the open star $\mathrm{st}(v)$ of a vertex $v \in K$ is given by

$$\mathrm{st}(v) = \{x \in |K| \mid t_v(x) = 1\},$$

and it is an open subset of $|K|$ (the set $|K| - \mathrm{st}(v)$ is the union of the polytopes of K that do not contain v as a vertex, a closed subset of $|K|$). Also, for any face σ, of K, the open star $\mathrm{st}(\sigma)$ of σ is given by

$$\mathrm{st}(\sigma) = \{x \in |K| \mid t_v(x) = 1, \text{ for all } v \in \mathrm{vert}(\sigma)\} = \bigcap_{v \in \mathrm{vert}(\sigma)} \mathrm{st}(v).$$

Therefore, $\mathrm{st}(\sigma)$ is also open in $|K|$.

The next proposition is another result that seems quite obvious, yet a rigorous proof is more involved that we might think. In fact, the only place that I am aware of where a proof is mentioned is the survey article by Carl Lee, Subdivisions and Triangulations of Polytopes (Chapter 17), in Goodman and O'Rourke [33]. Actually, the "proof" that Lee is referring to is a proof sketch whose details are "left to the reader." It turns out that a proof can be given using an inductive construction described in Grünbaum [36] (Chapter 5).

The proposition below 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.

One should be warned that even though, in the plane, every bounded region (not necessarily convex) whose boundary consists of a finite number of closed polygons (polygons homeomorphic to the circle S^1) can be triangulated, this is no longer true in three dimensions! For example, the 3D polyhedron known as the *Schönhart polyhedron* cannot be triangulated; see Boissonnat and Yvinec [12] (Chapters 13, Section 13.2).

Proposition 10.15. *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.*

Proof sketch. It would be tempting to proceed by induction on the dimension, d, of P but we do not know any correct proof of this kind. Instead, we proceed by induction on the number, p, of vertices of P. Since $\dim(P) = d$, we must have $p \geq d + 1$. The case $p = d + 1$ corresponds to a simplex, so the base case holds.

For $p > d + 1$, we can pick some vertex, $v \in P$, such that the convex hull, Q, of the remaining $p - 1$ vertices still has dimension d. Then, by the induction hypothesis, Q, has a simplicial subdivision. Now, we say that a facet, F, of Q is *visible from* v iff v and the interior of Q are strictly separated by the supporting hyperplane of F. Then, we add the d-simplices, $\mathrm{conv}(F \cup \{v\}) = v * F$, for every facet, F, of Q visible from v to those in the triangulation of Q. We claim that the resulting collection of simplices (with their faces) constitutes a simplicial complex subdividing P.

This is the part of the proof that requires some details. We say that v *is beneath a facet* F *of* Q iff v belongs to the open half–space determined by the supporting hyperplane of F which contains the interior of Q. We make use of a theorem of Grünbaum [36] (Theorem 1, Chapter 5, Section 5.2) which states the following:

Theorem (Grünbaum). *If P and Q are two polytopes as above with $P = \text{conv}(Q \cup \{v\})$, then the following properties hold:*

(i) A face F of Q is a face of P iff there is a facet F' of Q such that $F \subseteq F'$ and v is beneath F'.

(ii) If F is a face of Q, then $F^* = \text{conv}(F \cup \{v\})$ is a face of P iff either

(a) $v \in \text{aff}(F)$; or

(b) among the facets of Q containing F there is at least one such that v is beneath it, and at least one which is visible from v.

Moreover, each face of P is of one and only one of those types.

The above theorem implies that the new simplices that need to be added to form a triangulation of P are the convex hulls $\text{conv}(F \cup \{v\})$ associated with facets F of Q visible from v. The reader should check that everything really works out! □

With all this preparation, it is now quite natural to define combinatorial manifolds.

10.4 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 of its highly computational nature. This motivates the following definition:

Definition 10.11. 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* of *PL-manifold* for what we call a combinatorial 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 10.14 immediately yields the following result:

Proposition 10.16. *If X is a combinatorial d-manifold with boundary, then* $\text{bd}(X)$ *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 $\{U_i\}_{i\geq 0}$ of open sets of X 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 definition below.

Recall that

$$\mathbb{H}^d = \{(x_1, \ldots, x_d) \in \mathbb{R}^d \mid x_d \geq 0\}.$$

Definition 10.12. 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.

If $\varphi_i \colon U_i \to \Omega_i$ is some homeomorphism onto some open set Ω_i of $\mathbb{H}^d$ in the subset topology, some $p \in U_i$ may be mapped into $\mathbb{R}^{d-1} \times \mathbb{R}_+$, or into the "boundary" $\mathbb{R}^{d-1} \times \{0\}$ of $\mathbb{H}^d$. 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 a d-manifold, and that ∂M is closed. If $p \in \partial M$, and φ is some coordinate map defined on p, since $\Omega = \varphi(U)$ is an open subset of $\partial\mathbb{H}^d$, there is some open half ball B^d_{o+} centered at $\varphi(p)$ and contained in Ω which intersects $\partial\mathbb{H}^d$ along an open ball B^{d-1}_o, and if we consider $W = \varphi^{-1}(B^d_{o+})$, we have an open subset of M containing p which is mapped homeomorphically onto B^d_{o+} in such that way that every point in $W \cap \partial M$ is mapped onto the open ball B^{d-1}_o. Thus, it is easy to see that ∂M is a $(d-1)$-manifold.

Proposition 10.17. *Every combinatorial d-manifold is a d-manifold with boundary.*

Proof. This is an immediate consequence of Proposition 10.11. □

Is the converse of Proposition 10.17 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 is a manifold and that $M \cong |K|$ for some pure d-dimensional complex K?

Surprisingly, 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}). Such an example is the double suspension of Poincaré space; see Thurston [63], Example 3.2.11.

Fortunately, we will only have to deal with $d = 2, 3$! Another issue that must be addressed is orientability.

Assume that we 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:

$$\{(0, 1, 2), (1, 2, 0), (2, 0, 1)\} \quad \text{and} \quad \{(2, 1, 0), (1, 0, 2), (0, 2, 1)\}.$$

Definition 10.13. Let $X \cong |K|$ 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*.

Remark: It is possible to define the notion of orientation of a manifold but this is quite technical and we prefer to avoid digressing into this matter. This shows another advantage of combinatorial manifolds: The definition of orientability is simple and quite natural.

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(-Poincaré) 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 11

Shellings, the Euler–Poincaré Formula for Polytopes, the Dehn-Sommerville Equations and the Upper Bound Theorem

11.1 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. For this, Poincaré had to lay the foundations of algebraic topology and after a first "proof" given in 1893 (containing some flaws) he finally gave the first correct proof in 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 is called *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 [17] and soon after that (also in 1970) a striking application of shellability was made by McMullen [43] 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. This Chapter is heavily inspired by Ziegler's excellent treatment [69], Chapter 8. We begin with the definition of shellability. It's a bit technical, so please be patient!

Definition 11.1. Let K be a pure polyhedral complex of dimension d. A *shelling of* K is a list $F_1, \ldots, F_s$ of all 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, every set of points is shellable, by definition.

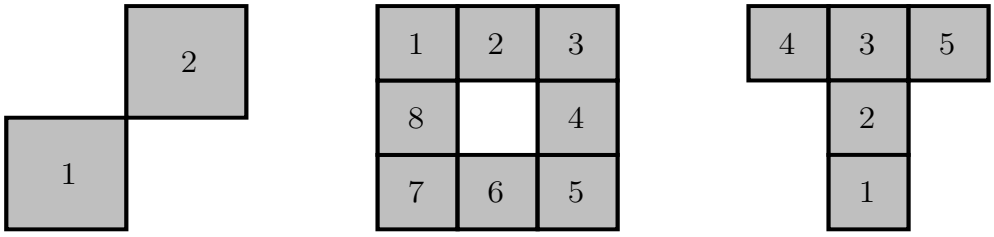


Figure 11.1: Non shellable and Shellable 2-complexes

(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.

However, already for 2-complexes, problems arise. For example, in Figure 11.1, 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. In the middle complex shown in Figure 11.1, the intersection of cell 8 with its predecessors is not connected, so the particular order chosen is not a shelling. However, there are other orders that constitute a shelling. 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 11.1 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; see Ziegler [69], Chapter 8.

Proposition 11.1. *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."

We have the following useful property of shellings of polytopes whose proof is left as an exercise (use induction on the dimension):

Proposition 11.2. *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 11.2 generally fails for complexes that are not polytopes, see the right 2-complex in Figure 11.1.

We will now present the proof that every polytope is shellable, using a technique invented by Bruggesser and Mani (1970) known as *line shelling* [17]. This is quite a simple and natural idea if one is willing to ignore the technical details involved in actually checking that it works. 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 11.2 (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. For a while, only face F_1 is visible, but when we reach the intersection z_2 of ℓ with the supporting line of F_2, the face F_2 becomes visible, and F_1 becomes invisible as it is now hidden by the supporting line of F_2. So far, we have seen the faces F_1 and F_2, *in that order*. As we continue traveling along ℓ, no new face becomes visible but for a more complicated polygon, other faces F_i would become visible one at a time as we reach the intersection z_i of ℓ with the supporting line of F_i, and the order in which these faces become visible corresponds to the ordering of the z_i's along the line ℓ. Then, 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. At this stage, the only visible face is F_3. We continue traveling upward on ℓ and we reach z_3, the intersection of the supporting line of F_3 with ℓ. At this moment, F_4 becomes visible, and F_3 disappears as it is now hidden by the supporting line of F_4. Note that F_5 is not visible at this stage. Finally, we reach z_4, the intersection of the supporting line of F_4 with ℓ, and at this moment the last facet F_5 becomes visible (and F_4 becomes invisible, F_3 being also invisible). 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.

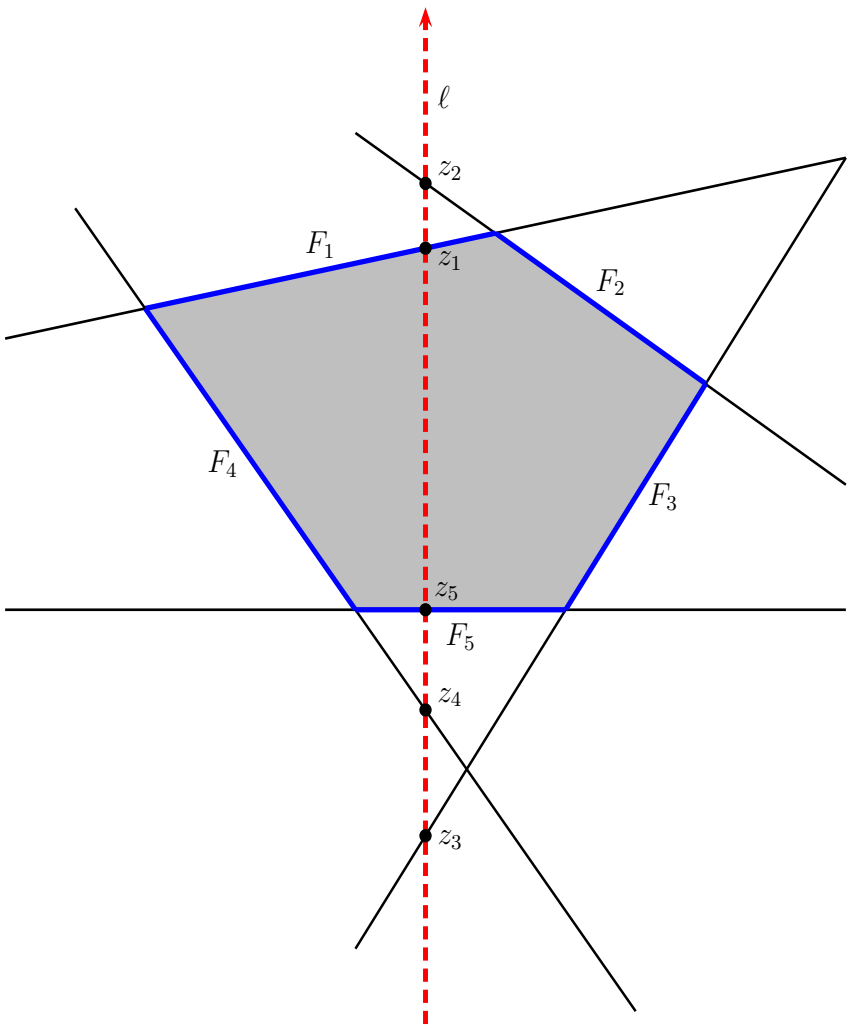


Figure 11.2: Shelling a polygon by travelling along a line

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 11.2. 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." The "trick" behind this particular perturbation method is that polynomials (in one variable) have a finite number of zeros.

Proposition 11.3. *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.

Proof. Assume that P is defined by t irredundant hyperplanes H_i, where H_i is given by the equation $a_i \cdot z = 1$, and write $\Lambda = (\lambda, \lambda^2, \ldots, \lambda^d)$ and $u = y - x$. Then the line ℓ_λ is given by

$$\ell_\lambda = \{x + s(y_\lambda - x) \mid s \in \mathbb{R}\} = \{x + s(u + \Lambda) \mid s \in \mathbb{R}\}.$$

The line ℓ_λ is not parallel to the hyperplane H_i iff

$$a_i \cdot (u + \Lambda) \neq 0, \quad i = 1, \ldots, t,$$

and it intersects the H_i in distinct points iff there is no $s \in \mathbb{R}$ such that

$$a_i \cdot (x + s(u + \Lambda)) = 1 \quad \text{and} \quad a_j \cdot (x + s(u + \Lambda)) = 1 \quad \text{for some } i \neq j.$$

Observe that $a_i \cdot (u + \Lambda) = p_i(\lambda)$ is a nonzero polynomial in λ of degree at most d. Since a polynomial of degree d has at most d zeros, if we let $Z(p_i)$ be the (finite) set of zeros of p_i we can ensure that ℓ_λ is not parallel to any of the H_i by picking $\lambda \notin \bigcup_{i=1}^t Z(p_i)$ (where $\bigcup_{i=1}^t Z(p_i)$ is a finite set). Now, as x is in general position w.r.t. P, we have $a_i \cdot x \neq 1$, for $i = 1 \ldots, t$. The condition stating that ℓ_λ intersects the H_i in distinct points can be written

$$a_i \cdot x + s a_i \cdot (u + \Lambda) = 1 \quad \text{and} \quad a_j \cdot x + s a_j \cdot (u + \Lambda) = 1 \quad \text{for some } i \neq j,$$

or

$$sp_i(\lambda) = \alpha_i \quad \text{and} \quad sp_j(\lambda) = \alpha_j \quad \text{for some } i \neq j,$$

where $\alpha_i = 1 - a_i \cdot x$ and $\alpha_j = 1 - a_j \cdot x$. As x is in general position w.r.t. P, we have $\alpha_i, \alpha_j \neq 0$ and as the H_i are irredundant, the polynomials $p_i(\lambda) = a_i \cdot (u + \Lambda)$ and $p_j(\lambda) = a_j \cdot (u + \Lambda)$ are not proportional. Now, if $\lambda \notin Z(p_i) \cup Z(p_j)$, in order for the system

$$\begin{aligned} sp_i(\lambda) &= \alpha_i \\ sp_j(\lambda) &= \alpha_j \end{aligned}$$

to have a solution in s we must have

$$q_{ij}(\lambda) = \alpha_i p_j(\lambda) - \alpha_j p_i(\lambda) = 0,$$

where $q_{ij}(\lambda)$ is not the zero polynomial since $p_i(\lambda)$ and $p_j(\lambda)$ are not proportional and $\alpha_i, \alpha_j \neq 0$. If we pick $\lambda \notin Z(q_{ij})$, then $q_{ij}(\lambda) \neq 0$. Therefore, if we pick

$$\lambda \notin \bigcup_{i=1}^t Z(p_i) \cup \bigcup_{i \neq j}^t Z(q_{ij}),$$

the line ℓ_λ is in general position w.r.t. P. Finally, we can pick λ small enough so that $y_\lambda = y + \Lambda$ is close enough to y so that it is in the interior of 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 11.3 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.

Finally, before proving the existence of line shellings for polytopes, we need more terminology.

Definition 11.3. 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 P only in y (equivalently, x and the interior of P are strictly separared by the supporting hyperplane of F).

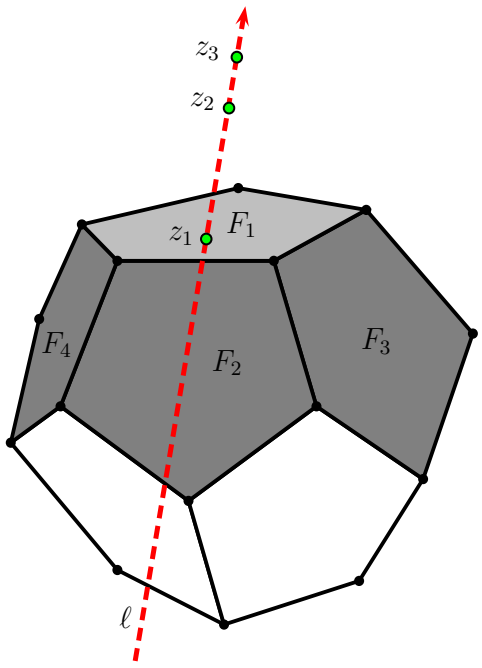


Figure 11.3: Shelling a polytope by travelling along a line, ℓ

We now prove the following fundamental theorem due to Bruggesser and Mani [17] (1970):

Theorem 11.4. *(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.*

Proof. By Proposition 11.3, we can find a line ℓ through x such that ℓ is in general position w.r.t. P and ℓ intersects the interior of P. Pick one of the two faces in which ℓ intersects P, say F_1, let $z_1 = \ell \cap F_1$, and orient ℓ from the inside of P to z_1. As ℓ intersects the supporting hyperplanes of the facets of P in distinct points, we get a linearly ordered list of these intersection points along ℓ,

$$z_1,\ z_2,\ \cdots,\ z_m,\ z_{m+1}, \cdots,\ z_s,$$

where z_{m+1} is the smallest element, z_m is the largest element, and where z_1 and z_s belong to the faces of P where ℓ intersects P. Then, as in the example illustrated by Figure 11.2, by travelling "upward" along the line ℓ starting from z_1 we get a total ordering of the facets of P,

$$F_1, F_2, \ldots, F_m, F_{m+1}, \ldots, F_s$$

where F_i is the facet whose supporting hyperplane cuts ℓ in z_i.

We claim that the above sequence is a shelling of P. This is proved by induction on d. For $d = 1$, P consists a line segment and the theorem clearly holds.

Consider the intersection $\partial F_j \cap (F_1 \cup \cdots \cup F_{j-1})$. We need to show that this is an initial segment of a shelling of ∂F_j. If $j \leq m$, *i.e.*, if F_j become visible before we reach ∞, then the above intersection is exactly the set of facets of F_j that are visible from $z_j = \ell \cap \mathrm{aff}(F_j)$. Therefore, by induction on the dimension, these facets are shellable and they form an initial segment of a shelling of the whole boundary ∂F_j.

If $j \geq m+1$, that is, after "passing through ∞" and reentering from $-\infty$, the intersection $\partial F_j \cap (F_1 \cup \cdots \cup F_{j-1})$ is the set of non-visible facets. By reversing the orientation of the line, ℓ, we see that the facets of this intersection are shellable and we get the reversed ordering of the facets.

Finally, when we reach the point x starting from z_1, the facets visible from x form an initial segment of the shelling, as claimed. □

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 [69] (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 11.5. *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.

Proof. For (1), we use a line in general position and intersecting F and F' in their interior. For (2), we pick a point x beyond v and pick a line in general position through x intersecting the interior of P. Pick the origin O in the interior of P. A point x is *beyond* v iff x and O lies on different sides of every hyperplane H_i supporting a facet of P containing x but on the same side of H_i for every hyperplane H_i supporting a facet of P **not** containing x. Such a point can be found on a line through O and v, as the reader should check. □

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 [25], Chapter 3). This result is an immediate

consequence of Corollary 11.5 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.

11.2 The Euler–Poincaré Formula for Polytopes

We begin by defining a very important topological concept, the Euler–Poincaré characteristic of a complex.

Definition 11.4. Let K be a d-dimensional polyhedral 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 [69] (Chapter 8), define the *reduced Euler–Poincaré characteristic* $\chi'(K)$ of a polyhedral complex (or a polytope) K as

$$\chi'(K) = -f_{-1} + f_0 - f_1 + f_2 + \cdots + (-1)^d f_d = \sum_{i=-1}^{d} (-1)^i f_i = -1 + \chi(K),$$

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, which 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

$$\chi(K_1 \cup K_2) = \chi(K_1) + \chi(K_2) - \chi(K_1 \cap K_2). \tag{$*$}$$

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 11.6. *(Euler–Poincaré Formula) For every polytope P of dimension d, 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 $(d, j) \geq (0, 1)$. For $d = 0$ and $j = 1$, the polytope P consists of a single point and so, $\chi(P) = f_0 = 1$, as claimed.

For the induction step, assume that $d \geq 1$. For $1 = j < f_{d-1}$, since F_1 is a polytope of dimension $d - 1$, by the induction hypothesis, $\chi(F_1) = 1$, as desired.

For $1 < j < f_{d-1}$, by $(*)$ we have

$$\chi(F_1 \cup \cdots F_{j-1} \cup F_j) = \chi\left(\bigcup_{i=1}^{j-1} F_i\right) + \chi(F_j) - \chi\left(\left(\bigcup_{i=1}^{j-1} F_i\right) \cap F_j\right). \qquad (**)$$

Since $(d, j-1) < (d, j)$, by the induction hypothesis,

$$\chi\left(\bigcup_{i=1}^{j-1} F_i\right) = 1$$

and since $\dim(F_j) = d - 1$, again by the induction hypothesis,

$$\chi(F_j) = 1.$$

Now, as $F_1, \ldots, F_{f_{d-1}}$ is a shelling and $j < f_{d-1}$, we have

$$\left(\bigcup_{i=1}^{j-1} F_i\right) \cap F_j = G_1 \cup \cdots \cup G_r,$$

for some shelling $G_1, \ldots, G_r, \ldots, G_t$ of $\mathcal{K}(\partial F_j)$, with $r < t = f_{d-2}(\partial F_j)$. The fact that $r < f_{d-2}(\partial F_j)$, *i.e.*, that $G_1 \cup \cdots \cup G_r$ is not the whole boundary of F_j is a property of line shellings and also follows from Proposition 11.2. As $\dim(\partial F_j) = d - 2$, and $r < f_{d-2}(\partial F_j)$, by the induction hypothesis, we have

$$\chi\left(\left(\bigcup_{i=1}^{j-1} F_i\right) \cap F_j\right) = \chi(G_1 \cup \cdots \cup G_r) = 1.$$

Consequently, by $(**)$ we obtain

$$\chi(F_1 \cup \cdots F_{j-1} \cup F_j) = 1 + 1 - 1 = 1,$$

as claimed (when $j < f_{d-1}$).

If $j = f_{d-1}$, then we have a complete shelling of $\partial F_{f_{d-1}}$, that is,

$$\left(\bigcup_{i=1}^{f_{d-1}-1} F_i\right) \cap F_{f_{d-1}} = G_1 \cup \cdots \cup G_{f_{d-2}(F_{f_{d-1}})} = \partial F_{f_{d-1}}.$$

As $\dim(\partial F_j) = d-2$, by the induction hypothesis,

$$\chi(\partial F_{f_{d-1}}) = \chi(G_1 \cup \cdots \cup G_{f_{d-2}(F_{f_{d-1}})}) = 1 - (-1)^{d-1},$$

and by $(**)$ it follows that

$$\chi(F_1 \cup \cdots \cup F_{f_{d-1}}) = 1 + 1 - (1 - (-1)^{d-1}) = 1 + (-1)^{d-1} = 1 - (-1)^d,$$

establishing the induction hypothesis in this last case. But then,

$$\chi(\partial P) = \chi(F_1 \cup \cdots \cup F_{f_{d-1}}) = 1 - (-1)^d$$

and

$$\chi(P) = \chi(\partial P) + (-1)^d = 1,$$

proving our theorem. □

Remark: Other combinatorial proofs of the Euler–Poincaré formula are given in Grünbaum [36] (Chapter 8), Boissonnat and Yvinec [12] (Chapter 7) and Ewald [26] (Chapter III). Coxeter gives a proof very close to Poincaré's own proof using notions of homology theory [20] (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!

11.3 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 [36] (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.

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 $i+1$).

In the rest of this section, we restrict our attention to pure 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 $F_1, \ldots, F_{j-1}$.

Definition 11.5. For any shelling $F_1, \ldots, F_s$ of a pure simplicial complex K of dimension $d-1$, for every j, with $1 \leq j \leq s$, the *restriction* R_j of the facet F_j is the set of "obligatory" vertices

$$R_j = \{v \in F_j \mid F_j - \{v\} \subseteq F_i, \text{ for some } i \text{ with } 1 \leq i < j\}.$$

Observe that $R_1 = \emptyset$. 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

$$I_j = \{G \subseteq V \mid R_j \subseteq G \subseteq F_j\}.$$

The proof of the above fact is left as an exercise to the reader, or see Ziegler [69] (Chapter 8, Section 8.3).

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 (a simplicial 1-dimensional complex) from Ziegler [69] (Section 8.3) shown in Figure 11.4.

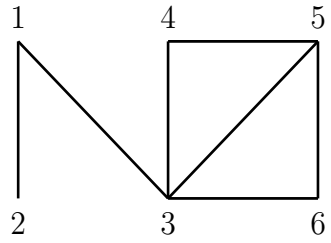


Figure 11.4: A connected 1-dimensional complex, $\mathcal{C}$.

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 $\mathcal{C}$ together with the blocks of the partition $\{I_1, \ldots, I_7\}$ associated with the seven edges of $\mathcal{C}$ are shown in Figure 11.5, with the blocks I_j shown in boldface.

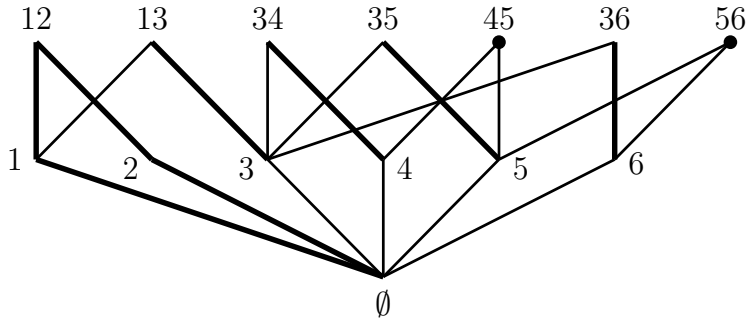


Figure 11.5: The partition associated with a shelling of $\mathcal{C}$.

In the above example, we have $R_1 = \{\emptyset\}$, $R_2 = \{3\}$, $R_3 = \{4\}$, $R_4 = \{5\}$, $R_5 = \{4, 5\}$, $R_6 = \{6\}$ and $R_7 = \{5, 6\}$, $I_1 = \{\emptyset, 1, 2, 12\}$, $I_2 = \{3, 13\}$, $I_3 = \{4, 34\}$, $I_4 = \{5, 35\}$, $I_5 = \{45\}$, $I_6 = \{6, 36\}$, $I_7 = \{56\}$, and the "minimal" new faces (corresponding to the R_j's) added at every stage of the shelling are

$$\emptyset,\ 3,\ 4,\ 5,\ 45,\ 6,\ 56.$$

Definition 11.6. For any shellable pure simplicial complex K of dimension $d-1$, if h_i is the number of blocks I_j such that the corresponding restriction set R_j has size i, that is,

$$h_i = |\{j \mid |R_j| = i,\ 1 \leq j \leq s\}| \quad \text{for } i = 0, \ldots, d,$$

then we define the *h-vector* associated with K as

$$\mathbf{h}(K) = (h_0, \ldots, h_d).$$

In other words, h_i is the number of minimal faces in the partitition that have i vertices, with $h_0 = 1$.

In our 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 have $h_0 = 1$, $h_1 = 4$, and $h_2 = 2$, that is,

$$\mathbf{h}(\mathcal{C}) = (1, 4, 2).$$

Looking at Figure 11.5, we see that for every horizontal layer i (starting from 0) of the lattice, h_i is the numbers of nodes (in bold) that are minimal in some block of the partition.

Now, let us show that if K is a shellable simplicial complex, then the f-vector can be recovered from the h-vector. Indeed, K is a pure simplicial complex so every face is contained in a face of dimension $d-1$ which has d vertices, and 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,

$$f_{k-1} = \sum_{j=1}^{s} \binom{d - |R_j|}{k - |R_j|},$$

and by definition of h_i, we get

$$f_{k-1} = \sum_{i=0}^{k} h_i \binom{d-i}{k-i} = h_k + \binom{d-k+1}{1} h_{k-1} + \cdots + \binom{d-1}{k-1} h_1 + \binom{d}{k} h_0, \qquad (*)$$

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

$$h_k = \sum_{i=0}^{k} (-1)^{k-i} \binom{d-i}{d-k} f_{i-1}.$$

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 11.7. For any $(d-1)$-dimensional pure 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

$$h_k = \sum_{i=0}^{k} (-1)^{k-i} \binom{d-i}{d-k} f_{i-1}.$$

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 [69] (Section 8.3).

We summarize below most of what we just showed:

Proposition 11.7. *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.*

There is a way of computing the h-vector of a pure simplicial complex from its f-vector reminiscent of the Pascal triangle (except that negative entries can turn up). This method is known as *Stanley's trick*; see Stanley [57]. For this we write the numbers f_i to the last entries of the rows of Pascal's triangle (to the place where ordinarily we would put $\binom{i+1}{i+1} = 1$), and then we compute the other entries using the rule

$$\text{upper right neighbor} \; - \; \text{upper left neighbor}.$$

For example, for the graph $\mathcal{C}$ of Figure 11.4, for which the f-vector is $\mathbf{f} = (1, 6, 7)$, we obtain the following table

$$\begin{array}{ccccc} & & \mathbf{1} & & \\ & 1 & & \mathbf{6} & \\ 1 & & 5 & & \mathbf{7} \\ 1 & & 4 & & 2 \end{array}$$

and we get $\mathbf{h}(\mathcal{C}) = (1, 4, 7)$.

If we now consider the boundary of an octahedron we have $\mathbf{f} = (1, 6, 12, 8)$, we have the following table

$$\begin{array}{ccccccc}
 & & & \mathbf{1} & & & \\
 & & 1 & & \mathbf{6} & & \\
 & 1 & & 5 & & \mathbf{12} & \\
1 & & 4 & & 7 & & \mathbf{8} \\
1 & & 3 & & 3 & & 1
\end{array}$$

so $\mathbf{h} = (1, 3, 3, 1)$.

For a simplicial complex that is not shellable, it is possible to obtain an h-vector with negative entries; see Ziegler [69] (Section 8.3, Example (iii)).

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 11.8. *(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,

$$f_{k-1} = \sum_{i=k}^{d} (-1)^{d-i} \binom{i}{k} f_{i-1} \qquad k = 0, \ldots, d.$$

Furthermore, the equation $h_0 = h_d$ is equivalent to the Euler–Poincaré formula.

Proof. We present a short and elegant proof due to McMullen. Recall from Proposition 11.2 that the reversal $F_s, \ldots, F_1$ of a shelling $F_1, \ldots, F_s$ of a polytope is also a shelling. From this, we see that for every F_j, the restriction set of F_j in the reversed shelling is equal to $R_j - F_j$, the complement of the restriction set of F_j in the original shelling. Therefore, if $|R_j| = k$, then F_j contributes "1" to h_k in the original shelling iff it contributes "1" to h_{d-k} in the reversed shelling (where $|R_j - F_j| = d - k$). It follows that the value of h_k computed in the

original shelling is the same as the value of h_{d-k} computed in the reversed shelling. However, by Proposition 11.7, the h-vector is independent of the shelling and hence, $h_k = h_{d-k}$.

To prove the second equation, following Ewald [26] (Chapter III, Theorem 3.7), define the polynomials $F(x)$ and $H(x)$ by

$$F(x) = \sum_{i=0}^{d} f_{i-1}x^i; \quad H(x) = (1-x)^d F\left(\frac{x}{1-x}\right).$$

Note that $H(x) = \sum_{i=0}^{d} f_{i-1}x^i(1-x)^{d-i}$, and an easy computation shows that the coefficient of x^k is equal to

$$\sum_{i=0}^{k}(-1)^{k-i}\binom{d-i}{d-k}f_{i-1} = h_k.$$

Now, the equations $h_k = h_{d-k}$ are equivalent to

$$H(x) = x^d H(x^{-1}). \tag{$\dagger$}$$

Substituting the definition of H in terms of F in equation ($\dagger$) yields

$$\begin{aligned}
(1-x)^d F\left(\frac{x}{1-x}\right) &= x^d(1-1/x)^d F\left(\frac{1/x}{1-1/x}\right)\\
(1-x)^d F\left(\frac{x}{1-x}\right) &= x^d\frac{(x-1)^d}{x^d}F\left(\frac{1}{x-1}\right)\\
(1-x)^d F\left(\frac{x}{1-x}\right) &= (-(1-x))^d F\left(\frac{1}{x-1}\right)\\
F\left(\frac{x}{1-x}\right) &= (-1)^d F\left(\frac{1}{x-1}\right),
\end{aligned}$$

where the last equation holds for all $x \neq 1$. If we let

$$y = -\frac{1}{x-1} = \frac{1}{1-x},$$

then

$$y - 1 = \frac{1}{1-x} - 1 = \frac{x}{1-x},$$

so the equation

$$F\left(\frac{x}{1-x}\right) = (-1)^d F\left(\frac{1}{x-1}\right)$$

which holds for all $x \neq 1$ yields

$$F(y-1) = (-1)^d F(-y)$$

for all $y \neq 0$ (since $y = 1/(1-x)$). But $F(x-1)$ and $(-1)^d F(-x)$ are polynomials that have the same value for infinitely many real values, so in fact the polynomials $F(x-1)$ and $(-1)^d F(-x)$ are identical. As

$$F(x-1) = \sum_{i=0}^{d} f_{i-1}(x-1)^i = \sum_{i=0}^{d} f_{i-1} \sum_{j=0}^{i} \binom{i}{i-j} x^{i-j}(-1)^j,$$

we see that the coefficient of x^k in $F(x-1)$ (obtained when $i-j=k$, that is, $j=i-k$) is

$$\sum_{i=0}^{d} (-1)^{i-k} \binom{i}{k} f_{i-1} = \sum_{i=k}^{d} (-1)^{i-k} \binom{i}{k} f_{i-1}.$$

On the other hand, the coefficient of x^k in $(-1)^d F(-x)$ is $(-1)^{d+k} f_{k-1}$. By equating the coefficients of x^k, we get

$$(-1)^{d+k} f_{k-1} = \sum_{i=k}^{d} (-1)^{i-k} \binom{i}{k} f_{i-1},$$

which, by multiplying both sides by $(-1)^{d+k}$, is equivalent to

$$f_{k-1} = \sum_{i=k}^{d} (-1)^{d+i} \binom{i}{k} f_{i-1} = \sum_{i=k}^{d} (-1)^{d-i} \binom{i}{k} f_{i-1},$$

as claimed. Finally, as we already know that

$$h_d = f_{d-1} - f_{d-2} + f_{d-3} + \cdots + (-1)^{d-1} f_0 + (-1)^d$$

and $h_0 = 1$, by multiplying both sides of the equation $h_d = h_0 = 1$ by $(-1)^{d-1}$ and moving $(-1)^d(-1)^{d-1} = -1$ to the right hand side, we get 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 [36], Chapter 9).

Remark: Readers familiar with homology and cohomology may suspect that the Dehn–Sommerville equations are a consequence of a type of Poincaré duality. Stanley proved that this is indeed the case. It turns out that the h_i are the dimensions of cohomology groups of a certain *toric variety* associated with the polytope. For more on this topic, see Stanley [58] (Chapters II and III) and Fulton [28] (Section 5.6).

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 [43]. We will present this result (without proof) in the next section.

11.4 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 [43] (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 [2] (1985). A version of this proof can also be found in Ewald [26] (Chapter 3).

McMullen's proof is not really very difficult but it is still quite involved so we will only state some propositions needed for its proof. We urge the reader to read Ziegler's account of this beautiful proof [69] (Chapter 8). We begin with cyclic polytopes.

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 11.8. 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 11.9. *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.*

Proof. We may assume that $n = d+1$. Say $c(t_1), \ldots, c(t_n)$ belong to a hyperplane, H, given by

$$\alpha_1 x_1 + \cdots + \alpha_d x_d = \beta.$$

(Of course, not all the α_i are zero.) Then, we have the polynomial, $H(t)$, given by

$$H(t) = -\beta + \alpha_1 t + \alpha_2 t^2 + \cdots + \alpha_d t^d,$$

of degree at most d and as each $c(t_i)$ belong to H, we see that each $c(t_i)$ is a zero of $H(t)$. However, there are $d+1$ distinct $c(t_i)$, so $H(t)$ would have $d+1$ distinct roots. As $H(t)$ has degree at most d, it must be the zero polynomial, a contradiction. Returing to the original $n > d+1$, we just proved every $d+1$ of the points $c(t_1), \ldots, c(t_n)$ are affinely independent. Then, every proper face of $C_d(n)$ has at most d independent vertices, which means that it is a simplex. □

The following proposition already shows that the cyclic polytope, $C_d(n)$, has $\binom{n}{k}$ $(k-1)$-faces if $1 \leq k \leq \lfloor \frac{d}{2} \rfloor$.

Proposition 11.10. *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*

$$f_k(C_d(n)) = \binom{n}{k+1} \qquad \text{if} \quad 0 \leq k < \left\lfloor \frac{d}{2} \right\rfloor.$$

Proof. Consider any sequence $t_{i_1} < t_{i_2} < \cdots < t_{i_k}$. We will prove that there is a hyperplane separating $F = \text{conv}(\{c(t_{i_1}), \ldots, c(t_{i_k})\})$ and $C_d(n)$. Consider the polynomial

$$p(t) = \prod_{j=1}^{k} (t - t_{i_j})^2$$

and write

$$p(t) = a_0 + a_1 t + \cdots + a_{2k} t^{2k}.$$

Consider the vector

$$a = (a_1, a_2, \ldots, a_{2k}, 0, \ldots, 0) \in \mathbb{R}^d$$

and the hyperplane, H, given by

$$H = \{x \in \mathbb{R}^d \mid x \cdot a = -a_0\}.$$

Then, for each j with $1 \leq j \leq k$, we have

$$c(t_{i_j}) \cdot a = a_1 t_{i_j} + \cdots + a_{2k} t_{i_j}^{2k} = p(t_{i_j}) - a_0 = -a_0,$$

and so, $c(t_{i_j}) \in H$. On the other hand, for any other point, $c(t_i)$, distinct from any of the $c(t_{i_j})$, we have

$$c(t_i) \cdot a = -a_0 + p(t_i) = -a_0 + \prod_{j=1}^{k} (t_i - t_{i_j})^2 > -a_0,$$

proving that $c(t_i) \in H_+$. But then, H is a supporting hyperplane of F for $C_d(n)$ and F is a $(k-1)$-face. □

Observe that Proposition 11.10 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 11.10 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 11.11. *(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*

$$C_d(n) = \text{conv}(c(t_1), \ldots, c(t_n)).$$

A subset $S \subseteq \{t_1, \ldots, t_n\}$ 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:

$$|\{k \in S \mid i < k < j\}| \equiv 0 \,(\text{mod } 2) \quad \textit{for} \quad i, j \notin S$$

Proof. Write $S = \{s_1, \ldots, s_d\} \subseteq \{t_1, \ldots, t_n\}$. Consider the polyomial

$$q(t) = \prod_{i=1}^{d}(t - s_i) = \sum_{j=0}^{d} b_j t^j,$$

let $b = (b_1, \ldots, b_d)$, and let H be the hyperplane given by

$$H = \{x \in \mathbb{R}^d \mid x \cdot b = -b_0\}.$$

Then, for each i, with $1 \leq i \leq d$, we have

$$c(s_i) \cdot b = \sum_{j=1}^{d} b_j s_i^j = q(s_i) - b_0 = -b_0,$$

so that $c(s_i) \in H$. For all other $t \neq s_i$,

$$q(t) = c(t) \cdot b + b_0 \neq 0,$$

that is, $c(t) \notin H$. Therefore, $F = \{c(s_1), \ldots, c(s_d)\}$ is a facet of $C_d(n)$ iff $\{c(t_1), \ldots, c(t_n)\} - F$ lies in one of the two open half-spaces determined by H. This is equivalent to $q(t)$ changing its sign an even number of times while, increasing t, we pass through the vertices in F. Therefore, the proposition is proved. □

In particular, Proposition 11.11 shows that the combinatorial structure of $C_d(n)$ does not depend on the specific choice of the sequence $t_1 < \cdots < t_n$. This justifies our notation $C_d(n)$.

Here is the celebrated upper bound theorem first proved by McMullen [43].

Theorem 11.12. *(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 11.12 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 11.13. *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.

Sketch of proof. First, we apply Proposition 10.15 to triangulate the facets of P without adding any vertices. Then, we can perturb the vertices to obtain a simplicial polytope P' with at least as many facets (and thus, faces) as P. □

Proposition 11.13 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 [69], 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 $k \leq \lfloor \frac{d}{2} \rfloor$, then the upper bound theorem will be proved. Indeed, McMullen proved the following theorem which is "more than enough" to yield the desired result ([43]):

Theorem 11.14. *(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.)

The proof of Theorem 11.14 is too involved to be given here, which is unfortunate since it is really beautiful. It makes a clever use of shellings and a careful analysis of the h-numbers of links of vertices. Again, the reader is referred to Ziegler [69], Chapter 8.

Since cyclic d-polytopes are neighborly (which means that they are $\lfloor \frac{d}{2} \rfloor$-neighborly), Theorem 11.12 follows from Proposition 11.13, and Theorem 11.14.

Corollary 11.15. *For every simplicial neighborly d-polytope with n vertices, we have*

$$f_{k-1} = \sum_{i=0}^{\lfloor \frac{d}{2} \rfloor *} \left(\binom{d-i}{k-i} + \binom{i}{k-d+i} \right) \binom{n-d-1+i}{i} \qquad \text{for} \quad 1 \leq k \leq d.$$

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 11.15 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 [54].

Consider any shelling of any simplicial d-polytope, P. For every facet, F_j, of a shelling either the restriction set R_j or its complement $F_j - R_j$ has at most $\lfloor \frac{d}{2} \rfloor$ elements. So, either in the shelling or in the reversed shelling, the restriction set of F_j has at most $\lfloor \frac{d}{2} \rfloor$ elements. Moreover, the restriction sets are all distinct, by construction. Thus, the number of facets is at most twice the number of k-faces of P with $k \leq \lfloor \frac{d}{2} \rfloor$. It follows that

$$f_{d-1} \leq 2 \sum_{i=0}^{\lfloor \frac{d}{2} \rfloor} \binom{n}{i}$$

and this rough estimate yields a $O(n^{\lfloor \frac{d}{2} \rfloor})$ bound.

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 [5] summarizes progress in this area up to 1993.

2. Billera and Björner [11] is a more advanced survey which reports on results up to 1997.

In fact, many of the chapters in Goodman and O'Rourke [33] should be of interest to the reader.

Generalizations of the Upper Bound Theorem using sophisticated techniques (face rings) due to Stanley can be found in Stanley [58] (Chapters II) and connections with toric varieties can be found in Stanley [58] (Chapters III) and Fulton [28].

Chapter 12

Projective Spaces, Projective Polyhedra, Polar Duality w.r.t. a Nondegenerate Quadric

The fact that not just points but also vectors are needed to deal with unbounded polyhedra is a hint that perhaps the notions of polytope and polyhedra can be unified by "going projective". Indeed, the goal of this chapter is to define a notion of projective polyhedron which is a natural extension of the notion of polyhedron in affine space, and retains many of the properties of polyhedra.

However, we have to be careful because projective geometry does not accommodate well the notion of convexity. This is because convexity has to do with convex combinations, but the essence of projective geometry is that everything is defined up to *non-zero* scalars, without any requirement that these scalars be positive.

It is possible to develop a theory of *oriented projective geometry* (due to J. Stolfi [59]) in which convexity is nicely accommodated. However, in this approach, every point comes as a pair, (positive point, negative point), and although it is a very elegant theory, we find it a bit unwieldy. However, since all we really need is to "embed" $\mathbb{E}^d$ into its *projective completion*, $\mathbb{P}^d$, so that we can deal with "points at infinity" and "normal points" in a uniform manner in particular, with respect to projective transformations, we will content ourselves with a definition of a notion of projective polyhedron using the notion of polyhedral cone. This notion is just what is needed in Chapter 13 to deal with the correspondence between Voronoi diagrams and Delaunay triangulations in terms of the lifting to a paraboloid or the lifting to a sphere. We will not attempt to define a general notion of convexity.

12.1 Projective Spaces

We begin with a "crash course" on (real) projective spaces. There are many texts on projective geometry. We suggest starting with Gallier [30] and then move on to far more

comprehensive treatments such as Berger (Geometry II) [8] or Samuel [52].

Definition 12.1. The *(real) projective space* $\mathbb{RP}^n$ is the set of all lines through the origin in $\mathbb{R}^{n+1}$, i.e., the set of one-dimensional subspaces of $\mathbb{R}^{n+1}$ (where $n \geq 0$). Since a one-dimensional subspace $L \subseteq \mathbb{R}^{n+1}$ is spanned by any nonzero vector $u \in L$, we can view $\mathbb{RP}^n$ as the set of equivalence classes of nonzero vectors in $\mathbb{R}^{n+1} - \{0\}$ modulo the equivalence relation,

$$u \sim v \quad \text{iff} \quad v = \lambda u, \quad \text{for some} \quad \lambda \in \mathbb{R},\ \lambda \neq 0.$$

We have the projection $p\colon (\mathbb{R}^{n+1} - \{0\}) \to \mathbb{RP}^n$ given by $p(u) = [u]_\sim$, the equivalence class of u modulo $\sim$. Write $[u]$ (or $\langle u \rangle$) for the line

$$[u] = \{\lambda u \mid \lambda \in \mathbb{R}\}$$

defined by the nonzero vector u. Note that $[u]_\sim = [u] - \{0\}$ for every $u \neq 0$, so the map $[u]_\sim \mapsto [u]$ is a bijection which allows us to identify $[u]_\sim$ and $[u]$. Thus, we will use both notations interchangeably as convenient.

The projective space $\mathbb{RP}^n$ is sometimes denoted $\mathbb{P}(\mathbb{R}^{n+1})$. Since every line L in $\mathbb{R}^{n+1}$ intersects the sphere S^n in two antipodal points, we can view $\mathbb{RP}^n$ as the quotient of the sphere S^n by identification of antipodal points. We call this the *spherical model* of $\mathbb{RP}^n$, which we illustrate in Figure 12.1.

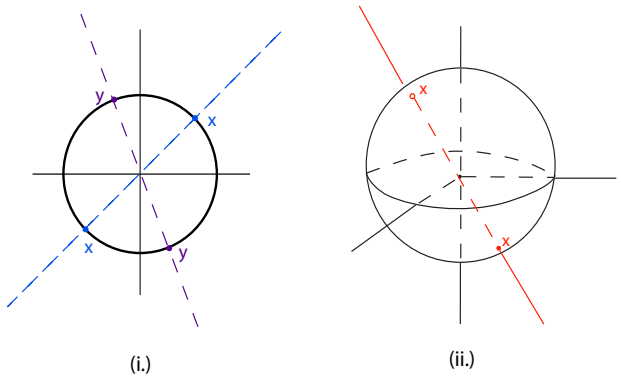


Figure 12.1: The geometric construction for $\mathbb{RP}^1$ and $\mathbb{RP}^2$ via the identification of antipodal points of S^1 and S^2 respectively.

A more subtle construction consists in considering the (upper) half-sphere instead of the sphere, where the upper half-sphere S^n_+ is set of points on the sphere S^n such that $x_{n+1} \geq 0$. This time, every line through the center intersects the (upper) half-sphere in a single point, except on the boundary of the half-sphere, where it intersects in two antipodal points a_+

and a_-. Thus, the projective space $\mathbb{RP}^n$ is the quotient space obtained from the (upper) half-sphere S^n_+ by identifying antipodal points a_+ and a_- on the boundary of the half-sphere. We call this model of $\mathbb{RP}^n$ the *half-spherical model*, which we illustrate in Figure 12.2.

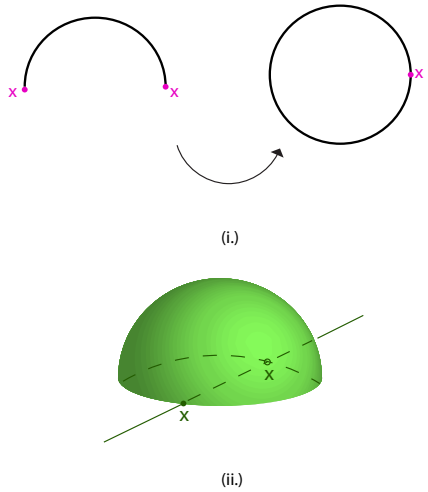


Figure 12.2: The geometric construction for $\mathbb{RP}^1 \sim S^1$ and $\mathbb{RP}^2$ in terms of the antipodal boundary points of S^1_+ and S^2_+ respectively.

When $n = 2$, we get a circle. When $n = 3$, the upper half-sphere is homeomorphic to a closed disk (say, by orthogonal projection onto the xy-plane), and $\mathbb{RP}^2$ is in bijection with a closed disk in which antipodal points on its boundary (a unit circle) have been identified. This is hard to visualize! In this model of the real projective space, projective lines are great semicircles on the upper half-sphere, with antipodal points on the boundary identified. Boundary points correspond to points at infinity. By orthogonal projection, these great semicircles correspond to semiellipses, with antipodal points on the boundary identified. Traveling along such a projective "line," when we reach a boundary point, we "wrap around"! In general, the upper half-sphere S^n_+ is homeomorphic to the closed unit ball in $\mathbb{R}^n$, whose boundary is the $(n-1)$-sphere S^{n-1}. For example, the projective space $\mathbb{RP}^3$ is in bijection with the closed unit ball in $\mathbb{R}^3$, with antipodal points on its boundary (the sphere S^2) identified!

Another useful way of "visualizing" $\mathbb{RP}^n$ is to use the hyperplane $H_{n+1} \subseteq \mathbb{R}^{n+1}$ of equation $x_{n+1} = 1$. Observe that for every line $[u]$ through the origin in $\mathbb{R}^{n+1}$, if u does not belong to the hyperplan $H_{n+1}(0) \cong \mathbb{R}^n$ of equation $x_{n+1} = 0$, then $[u]$ intersects H_{n+1} is a unique point, namely

$$\left(\frac{u_1}{u_{n+1}}, \ldots, \frac{u_n}{u_{n+1}}, 1\right),$$

where $u = (u_1, \ldots, u_{n+1})$. The lines $[u]$ for which $u_{n+1} = 0$ are "points at infinity". See Figure 12.3.

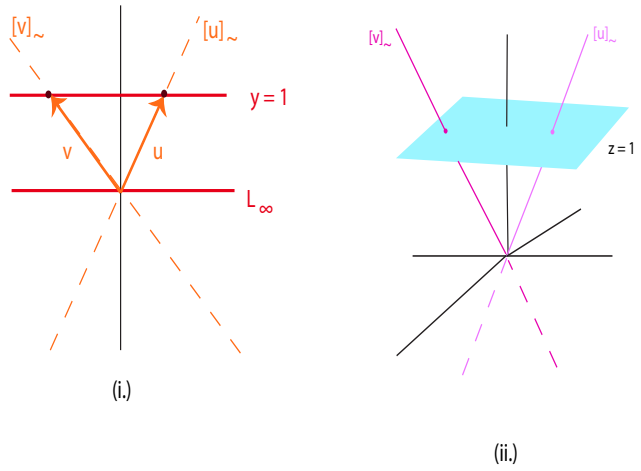


Figure 12.3: The plane model construction for $\mathbb{RP}^1$ and $\mathbb{RP}^2$, where points at infinity correspond to the x-axis and the xy-plane respectively.

Observe that the set of lines in $H_{n+1}(0) \cong \mathbb{R}^n$ is the set of points of the projective space $\mathbb{RP}^{n-1}$, and so $\mathbb{RP}^n$ can be written as the disjoint union

$$\mathbb{RP}^n = \mathbb{R}^n \amalg \mathbb{RP}^{n-1}.$$

We can repeat the above analysis on $\mathbb{RP}^{n-1}$ and so we can think of $\mathbb{RP}^n$ as the disjoint union

$$\mathbb{RP}^n = \mathbb{R}^n \amalg \mathbb{R}^{n-1} \amalg \cdots \amalg \mathbb{R}^1 \amalg \mathbb{R}^0,$$

where $\mathbb{R}^0 = \{0\}$ consist of a single point. The above shows that there is an embedding $\mathbb{R}^n \hookrightarrow \mathbb{RP}^n$ given by $(u_1, \ldots, u_n) \mapsto (u_1, \ldots, u_n, 1)$.

It will also be very useful to use homogeneous coordinates.

Definition 12.2. Given any point, $a = [u]_\sim \in \mathbb{RP}^n$, the set

$$\{(\lambda u_1, \ldots, \lambda u_{n+1}) \mid \lambda \neq 0\}$$

is called the set of *homogeneous coordinates* of a. Since $u \neq 0$, observe that for all homogeneous coordinates, $(u_1, \ldots, u_{n+1})$, for a, some u_i must be non-zero. The traditional notation for the homogeneous coordinates of a point $a = [u]_\sim$ is

$$(u_1 : \cdots : u_n : u_{n+1}).$$

There is a useful bijection between certain kinds of subsets of $\mathbb{R}^{d+1}$ and subsets of $\mathbb{RP}^d$. For any subset S of $\mathbb{R}^{d+1}$, let

$$-S = \{-u \mid u \in S\}.$$

Geometrically, $-S$ is the reflexion of S about 0. Note that for *any* nonempty subset, $S \subseteq \mathbb{R}^{d+1}$, with $S \neq \{0\}$, the sets S, $-S$, and $S \cup -S$ all induce the *same* set of points in projective space $\mathbb{RP}^d$, since

$$\begin{aligned} p(S - \{0\}) &= \{[u]_\sim \mid u \in S - \{0\}\} \\ &= \{[-u]_\sim \mid u \in S - \{0\}\} \\ &= \{[u]_\sim \mid u \in -S - \{0\}\} = p((-S) - \{0\}) \\ &= x\{[u]_\sim \mid u \in S - \{0\}\} \cup \{[u]_\sim \mid u \in (-S) - \{0\}\} = p((S \cup -S) - \{0\}), \end{aligned}$$

because $[u]_\sim = [-u]_\sim$. Using these facts we obtain a bijection between subsets of $\mathbb{RP}^d$ and certain subsets of $\mathbb{R}^{d+1}$.

Definition 12.3. We say that a set $S \subseteq \mathbb{R}^{d+1}$ is *symmetric* iff $S = -S$. Obviously, $S \cup -S$ is symmetric for any set S. Say that a subset $C \subseteq \mathbb{R}^{d+1}$ is a *double cone* iff for every $u \in C - \{0\}$, the entire line $[u]$ spanned by u is contained in C. See Figure 12.4.

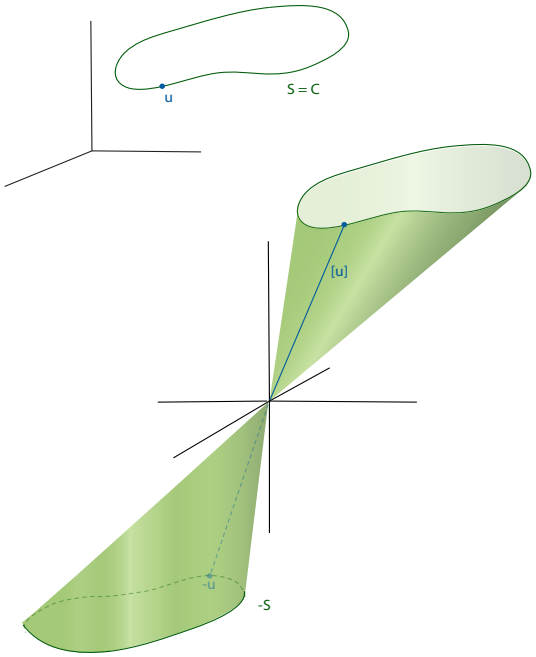


Figure 12.4: The double cone for the green curve S.

We exclude the trivial double cone, $C = \{0\}$, since the trivial vector space does not yield a projective space. Thus, every double cone can be viewed as a set of lines through 0. Note

that a double cone is symmetric. Given any nonempty subset, $S \subseteq \mathbb{RP}^d$, let $v(S) \subseteq \mathbb{R}^{d+1}$ be the set of vectors,

$$v(S) = \bigcup_{[u]_\sim \in S} [u]_\sim \cup \{0\}.$$

Note that $v(S)$ is a double cone.

Proposition 12.1. *The map, $v\colon S \mapsto v(S)$, from the set of nonempty subsets of $\mathbb{RP}^d$ to the set of nonempty, nontrivial double cones in $\mathbb{R}^{d+1}$ is a bijection.*

Proof. We already noted that $v(S)$ is nontrivial double cone. Consider the map,

$$\mathrm{ps}\colon S \mapsto p(S) = \{[u]_\sim \in \mathbb{RP}^d \mid u \in S - \{0\}\}.$$

We leave it as an easy exercise to check that $\mathrm{ps} \circ v = \mathrm{id}$ and $v \circ \mathrm{ps} = \mathrm{id}$, which shows that v and ps are mutual inverses. □

Definition 12.4. Given any subspace $X \subseteq \mathbb{R}^{n+1}$ with $\dim X = k+1 \geq 1$ and $0 \leq k \leq n$, a *k-dimensional projective subspace* of $\mathbb{RP}^n$ is the image $Y = p(X - \{0\})$ of $X - \{0\}$ under the projection p. We often write $Y = \mathbb{P}(X)$. When $k = n-1$, we say that Y is a *projective hyperplane* or simply a *hyperplane*. When $k = 1$, we say that Y is a *projective line* or simply a *line*. See Figure 12.5.

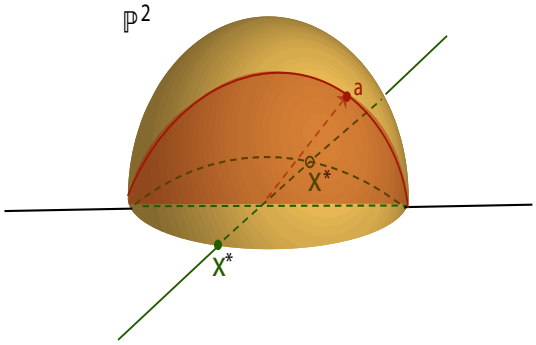


Figure 12.5: In the half-spherical model, a projective line is the maroon semi-circle obtained by intersecting the hemisphere with a plane through the origin.

It is easy to see that every projective hyperplane, H, is the kernel (zero set) of some linear equation of the form

$$a_1 x_1 + \cdots + a_{n+1} x_{n+1} = 0,$$

where one of the a_i is nonzero, in the sense that

$$H = \{(x_1 \colon \cdots \colon x_{n+1}) \in \mathbb{RP}^n \mid a_1 x_1 + \cdots + a_{n+1} x_{n+1} = 0\}.$$

Conversely, the kernel of any such linear equation defines a projective hyperplane. Furthermore, given a projective hyperplane, $H \subseteq \mathbb{RP}^n$, the linear equation defining H is unique up to a nonzero scalar.

Definition 12.5. For any i, with $1 \leq i \leq n+1$, the set

$$U_i = \{(x_1 \colon \cdots \colon x_{n+1}) \in \mathbb{RP}^n \mid x_i \neq 0\}$$

is a subset of $\mathbb{RP}^n$ called an *affine patch* of $\mathbb{RP}^n$.

We have a bijection, $\varphi_i \colon U_i \to \mathbb{R}^n$, between U_i and $\mathbb{R}^n$ given by

$$\varphi_i \colon (x_1 \colon \cdots \colon x_{n+1}) \mapsto \left(\frac{x_1}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_{n+1}}{x_i}\right).$$

This map is well defined because if $(y_1, \ldots, y_{n+1}) \sim (x_1, \ldots, x_{n+1})$, that is, $(y_1, \ldots, y_{n+1}) = \lambda(x_1, \ldots, x_{n+1})$, with $\lambda \neq 0$, then

$$\frac{y_j}{y_i} = \frac{\lambda x_j}{\lambda x_i} = \frac{x_j}{x_i} \qquad (1 \leq j \leq n+1),$$

since $\lambda \neq 0$ and $x_i, y_i \neq 0$. The inverse, $\psi_i \colon \mathbb{R}^n \to U_i \subseteq \mathbb{RP}^n$, of φ_i is given by

$$\psi_i \colon (x_1, \cdots, x_n) \mapsto (x_1 \colon \cdots x_{i-1} \colon 1 \colon x_i \colon \cdots \colon x_n).$$

Observe that the bijection, φ_i, between U_i and $\mathbb{R}^n$ can also be viewed as the bijection

$$(x_1 \colon \cdots \colon x_{n+1}) \mapsto \left(\frac{x_1}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, 1, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_{n+1}}{x_i}\right),$$

between U_i and the hyperplane, $H_i \subseteq \mathbb{R}^{n+1}$, of equation $x_i = 1$. We will make heavy use of these bijections. For example, for any subset, $S \subseteq \mathbb{RP}^n$, the "view of S from the patch U_i", $S \restriction U_i$, is in bijection with $v(S) \cap H_i$, where $v(S)$ is the double cone associated with S (see Proposition 12.1).

The affine patches, $U_1, \ldots, U_{n+1}$, cover the projective space $\mathbb{RP}^n$, in the sense that every $(x_1 \colon \cdots \colon x_{n+1}) \in \mathbb{RP}^n$ belongs to one of the U_i's, as not all $x_i = 0$. See Figures 12.6 and 12.7. The U_i's turn out to be open subsets of $\mathbb{RP}^n$ and they have nonempty overlaps. When we restrict ourselves to one of the U_i, we have an "affine view of $\mathbb{RP}^n$ from U_i." In particular, on the affine patch U_{n+1}, we have the "standard view" of $\mathbb{R}^n$ embedded into $\mathbb{RP}^n$ as H_{n+1}, the hyperplane of equation $x_{n+1} = 1$. The complement $H_i(0)$ of U_i in $\mathbb{RP}^n$ is the (projective) hyperplane of equation $x_i = 0$ (a copy of $\mathbb{RP}^{n-1}$). With respect to the affine patch U_i, the hyperplane $H_i(0)$ plays the role of *hyperplane (of points) at infinity*.

From now on, for simplicity of notation, we will write $\mathbb{P}^n$ for $\mathbb{RP}^n$. We need to define projective maps. Such maps are induced by linear maps.

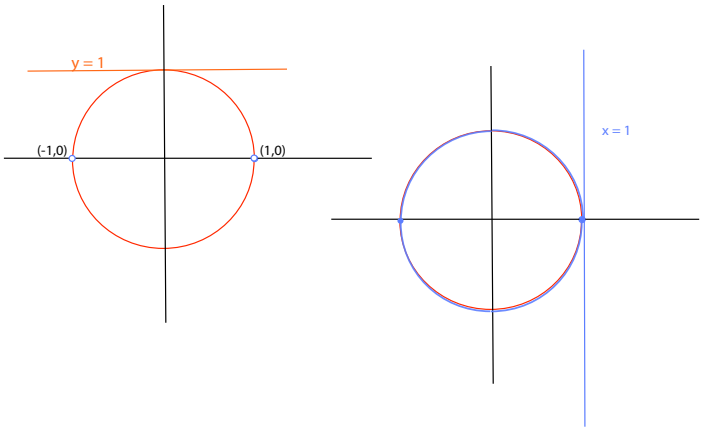


Figure 12.6: The space $\mathbb{RP}^1$, visualized by the spherical model, is covered by the two affine patches $y = 1$ or U_2, and $x = 1$ or U_1.

Definition 12.6. Any injective linear map, $h\colon \mathbb{R}^{m+1} \to \mathbb{R}^{n+1}$, induces a map, $\mathbb{P}(h)\colon \mathbb{P}^m \to \mathbb{P}^n$, defined by

$$\mathbb{P}(h)([u]_\sim) = [h(u)]_\sim$$

and called a *projective map*. When $m = n$ and h is bijective, the map $\mathbb{P}(h)$ is also bijective and it is called a *projectivity*.

We have to check that this definition makes sense, that is, it is compatible with the equivalence relation, $\sim$. For this, assume that $u \sim v$, that is

$$v = \lambda u,$$

with $\lambda \neq 0$ (of course, $u, v \neq 0$). As h is linear, we get

$$h(v) = h(\lambda u) = \lambda h(u),$$

that is, $h(u) \sim h(v)$, which shows that $[h(u)]_\sim$ does not depend on the representative chosen in the equivalence class of $[u]_\sim$. It is also easy to check that whenever two linear maps, h_1 and h_2, induce the same projective map, *i.e.*, if $\mathbb{P}(h_1) = \mathbb{P}(h_2)$, then there is a nonzero scalar, λ, so that $h_2 = \lambda h_1$.

Why did we require h to be injective? Because if h has a nontrivial kernel, then, any nonzero vector $u \in \mathrm{Ker}\,(h)$ is mapped to 0, but as 0 does **not** correspond to any point of $\mathbb{P}^n$, the map $\mathbb{P}(h)$ is undefined on $\mathbb{P}(\mathrm{Ker}\,(h))$.

In some case, we allow projective maps induced by non-injective linear maps h. In this case, $\mathbb{P}(h)$ is a map whose domain is $\mathbb{P}^n - \mathbb{P}(\mathrm{Ker}\,(h))$. An example is the map, $\sigma_N\colon \mathbb{P}^3 \to \mathbb{P}^2$, given by

$$(x_1\colon x_2\colon x_3\colon x_4) \mapsto (x_1\colon x_2\colon x_4 - x_3),$$

which is undefined at the point $(0\colon 0\colon 1\colon 1)$. This map is the "homogenization" of the central projection (from the north pole, $N = (0, 0, 1)$) from $\mathbb{E}^3$ onto $\mathbb{E}^2$.

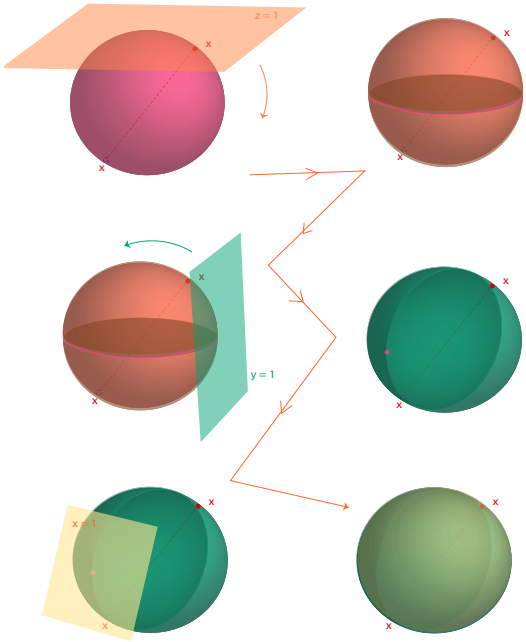


Figure 12.7: The space $\mathbb{RP}^2$, visualized by the spherical model, is covered by the three affine patches $z = 1$ or U_3, $y = 1$ or U_2, and $x = 1$ or U_1. The plane $z = 1$ covers everything except the pink circle $x^2 + y^2 = 1$. The plane $y = 1$ will cover this circle, excluding the x-intercepts. These x-intercepts are then covered by $x = 1$.

Although a projective map, $f\colon \mathbb{P}^m \to \mathbb{P}^n$, is induced by some linear map, h, the map f is **not** linear! This is because linear combinations of points in $\mathbb{P}^m$ *do not make any sense!*

Another way of defining functions (possibly partial) between projective spaces involves using homogeneous polynomials. If $p_1(x_1, \ldots, x_{m+1}), \ldots, p_{n+1}(x_1, \ldots, x_{m+1})$ are $n+1$ homogeneous polynomials all of the same degree d, and if these $n+1$ polynomials do not vanish simultaneously, then we claim that the function f given by

$$f(x_1\colon \cdots \colon x_{m+1}) = (p_1(x_1, \ldots, x_{m+1})\colon \cdots \colon p_{n+1}(x_1, \ldots, x_{m+1}))$$

is indeed a well-defined function from $\mathbb{P}^m$ to $\mathbb{P}^n$. Indeed, if $(y_1, \ldots, y_{m+1}) \sim (x_1, \ldots, x_{m+1})$, that is, $(y_1, \ldots, y_{m+1}) = \lambda(x_1, \ldots, x_{m+1})$, with $\lambda \neq 0$, as the p_i are homogeneous of degree d,

$$p_i(y_1, \ldots, y_{m+1}) = p_i(\lambda x_1, \ldots, \lambda x_{m+1}) = \lambda^d p_i(x_1, \ldots, x_{m+1}),$$

and so,

$$\begin{aligned} f(y_1\colon \cdots \colon y_{m+1}) &= (p_1(y_1,\ldots,y_{m+1})\colon \cdots \colon p_{n+1}(y_1,\ldots,y_{m+1})) \\ &= (\lambda^d p_1(x_1,\ldots,x_{m+1})\colon \cdots \colon \lambda^d p_{n+1}(x_1,\ldots,x_{m+1})) \\ &= \lambda^d (p_1(x_1,\ldots,x_{m+1})\colon \cdots \colon p_{n+1}(x_1,\ldots,x_{m+1})) \\ &= \lambda^d f(x_1\colon \cdots \colon x_{m+1}), \end{aligned}$$

which shows that $f(y_1\colon \cdots \colon y_{m+1}) \sim f(x_1\colon \cdots \colon x_{m+1})$, as required.

For example, the map, $\tau_N\colon \mathbb{P}^2 \to \mathbb{P}^3$, given by

$$(x_1\colon x_2, \colon x_3) \mapsto (2x_1x_3\colon 2x_2x_3\colon x_1^2+x_2^2-x_3^2\colon x_1^2+x_2^2+x_3^2),$$

is well-defined. It turns out to be the "homogenization" of the inverse stereographic map from $\mathbb{E}^2$ to S^2 (see Section 13.5). Observe that

$$\tau_N(x_1\colon x_2\colon 0) = (0\colon 0\colon x_1^2+x_2^2\colon x_1^2+x_2^2) = (0\colon 0\colon 1\colon 1),$$

that is, τ_N maps all the points at infinity (in $H_3(0)$) to the "north pole," $(0\colon 0\colon 1\colon 1)$. However, when $x_3 \neq 0$, we can prove that τ_N is injective (in fact, its inverse is σ_N, defined earlier).

Most interesting subsets of projective space arise as the collection of zeros of a (finite) set of homogeneous polynomials. Let us begin with a single homogeneous polynomial, $p(x_1,\ldots,x_{n+1})$, of degree d and set

$$V(p) = \{(x_1\colon \cdots \colon x_{n+1}) \in \mathbb{P}^n \mid p(x_1,\ldots,x_{n+1}) = 0\}.$$

As usual, we need to check that this definition does not depend on the specific representative chosen in the equivalence class of $[(x_1,\ldots,x_{n+1})]_\sim$. If $(y_1,\ldots,y_{n+1}) \sim (x_1,\ldots,x_{n+1})$, that is, $(y_1,\ldots,y_{n+1}) = \lambda(x_1,\ldots,x_{n+1})$, with $\lambda \neq 0$, as p is homogeneous of degree d,

$$p(y_1,\ldots,y_{n+1}) = p(\lambda x_1,\ldots,\lambda x_{n+1}) = \lambda^d p(x_1,\ldots,x_{n+1}),$$

and as $\lambda \neq 0$,

$$p(y_1,\ldots,y_{n+1}) = 0 \quad \text{iff} \quad p(x_1,\ldots,x_{n+1}) = 0,$$

which shows that $V(p)$ is well defined.

Definition 12.7. For a set of homogeneous polynomials (not necessarily of the same degree) $\mathcal{E} = \{p_1(x_1,\ldots,x_{n+1}),\ldots,p_s(x_1,\ldots,x_{n+1})\}$, we set

$$V(\mathcal{E}) = \bigcap_{i=1}^{s} V(p_i) = \{(x_1\colon \cdots \colon x_{n+1}) \in \mathbb{P}^n \mid p_i(x_1,\ldots,x_{n+1}) = 0,\ i = 1\ldots,s\}.$$

The set, $V(\mathcal{E})$, is usually called the *projective variety* defined by $\mathcal{E}$ (or *cut out by* $\mathcal{E}$). When $\mathcal{E}$ consists of a single polynomial p, the set $V(p)$ is called a (projective) *hypersurface*.

For example, if

$$p(x_1, x_2, x_3, x_4) = x_1^2 + x_2^2 + x_3^2 - x_4^2,$$

then $V(p)$ is the *projective sphere* in $\mathbb{P}^3$, denoted $\widetilde{S^2}$. Indeed, if we "look" at $V(p)$ on the affine patch U_4, where $x_4 \neq 0$, we know that this amounts to setting $x_4 = 1$, and we do get the set of points $(x_1, x_2, x_3, 1) \in U_4$ satisfying $x_1^2 + x_2^2 + x_3^2 - 1 = 0$, our usual 2-sphere! However, if we look at $V(p)$ on the patch U_1, where $x_1 \neq 0$, we see the quadric of equation $1 + x_2^2 + x_3^2 = x_4^2$, which is not a sphere but a hyperboloid of two sheets! Nevertheless, if we pick $x_4 = 0$ as the plane at infinity, note that the projective sphere does not have points at infinity since the only *real* solution of $x_1^2 + x_2^2 + x_3^2 = 0$ is $(0, 0, 0)$, but $(0, 0, 0, 0)$ does not correspond to any point of $\mathbb{P}^3$.

Another example is given by

$$q = (x_1, x_2, x_3, x_4) = x_1^2 + x_2^2 - x_3 x_4,$$

for which $V(q)$ corresponds to a paraboloid in the patch U_4. Indeed, if we set $x_4 = 1$, we get the set of points in U_4 satisfying $x_3 = x_1^2 + x_2^2$. For this reason, we denote $V(q)$ by $\widetilde{\mathcal{P}}$ and call it a *(projective) paraboloid*.

Definition 12.8. Given any homogeneous polynomial $F(x_1, \ldots, x_{d+1})$, we will also make use of the *hypersurface cone* $C(F) \subseteq \mathbb{R}^{d+1}$, defined by

$$C(F) = \{(x_1, \ldots, x_{d+1}) \in \mathbb{R}^{d+1} \mid F(x_1, \ldots, x_{d+1}) = 0\}.$$

Observe that $V(F) = \mathbb{P}(C(F))$.

Remark: Every variety $V(\mathcal{E})$, defined by a *set* of polynomials, $\mathcal{E} = \{p_1(x_1, \ldots, x_{n+1}), \ldots, p_s(x_1, \ldots, x_{n+1})\}$, is also the hypersurface defined by the *single* polynomial equation

$$p_1^2 + \cdots + p_s^2 = 0.$$

This fact, peculiar to the real field $\mathbb{R}$ is a mixed blessing. On the one-hand, the study of varieties is reduced to the study of hypersurfaces. On the other-hand, this is a hint that we should expect that such a study will be hard.

Perhaps to the surprise of the novice, there is a bijective projective map (a projectivity) sending $\widetilde{S^2}$ to $\widetilde{\mathcal{P}}$. This map, θ, is given by

$$\theta(x_1 \colon x_2 \colon x_3 \colon x_4) = (x_1 \colon x_2 \colon x_3 + x_4 \colon x_4 - x_3),$$

whose inverse is given by

$$\theta^{-1}(x_1 \colon x_2 \colon x_3 \colon x_4) = \left(x_1 \colon x_2 \colon \frac{x_3 - x_4}{2} \colon \frac{x_3 + x_4}{2}\right).$$

Indeed, if $(x_1 \colon x_2 \colon x_3 \colon x_4)$ satisfies

$$x_1^2 + x_2^2 + x_3^2 - x_4^2 = 0,$$

and if $(z_1 \colon z_2 \colon z_3 \colon z_4) = \theta(x_1 \colon x_2 \colon x_3 \colon x_4)$, then from above,

$$(x_1 \colon x_2 \colon x_3 \colon x_4) = \left(z_1 \colon z_2 \colon \frac{z_3 - z_4}{2} \colon \frac{z_3 + z_4}{2}\right),$$

and by plugging the right-hand sides in the equation of the sphere, we get

$$\begin{aligned} z_1^2 + z_2^2 + \left(\frac{z_3 - z_4}{2}\right)^2 - \left(\frac{z_3 + z_4}{2}\right)^2 &= z_1^2 + z_2^2 + \frac{1}{4}(z_3^2 + z_4^2 - 2z_3z_4 - (z_3^2 + z_4^2 + 2z_3z_4)) \\ &= z_1^2 + z_2^2 - z_3z_4 = 0, \end{aligned}$$

which is the equation of the paraboloid $\widetilde{\mathcal{P}}$.

12.2 Projective Polyhedra

Following the "projective doctrine" which consists in replacing points by lines through the origin, that is, to "conify" everything, we will define a projective polyhedron as any set of points in $\mathbb{P}^d$ induced by a polyhedral cone in $\mathbb{R}^{d+1}$. To do so, it is preferable to consider cones as sets of positive combinations of vectors (see Definition 5.3). Just to refresh our memory, a set, $C \subseteq \mathbb{R}^d$, 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, $u_1, \ldots, u_p \in \mathbb{R}^d$. An *$\mathcal{H}$-cone* is any subset of $\mathbb{R}^d$ given by a finite intersection of closed half-spaces cut out by hyperplanes through 0.

A good place to learn about cones (and much more) is Fulton [28]. See also Ewald [26].

By Theorem 5.19, $\mathcal{V}$-cones and $\mathcal{H}$-cones form the same collection of convex sets (for every $d \geq 0$). Naturally, we can think of these cones as sets of rays (half-lines) of the form

$$\langle u \rangle_+ = \{\lambda u \mid \lambda \in \mathbb{R},\ \lambda \geq 0\},$$

where $u \in \mathbb{R}^d$ is any *nonzero* vector. We exclude the trivial cone, $\{0\}$, since 0 does not define any point in projective space. When we "go projective," each ray corresponds to the full line, $\langle u \rangle$, spanned by u which can be expressed as

$$\langle u \rangle = \langle u \rangle_+ \cup -\langle u \rangle_+,$$

where $-\langle u \rangle_+ = \langle u \rangle_- = \{\lambda u \mid \lambda \in \mathbb{R},\ \lambda \leq 0\}$. Now, if $C \subseteq \mathbb{R}^d$ is a polyhedral cone, obviously $-C$ is also a polyhedral cone and the set $C \cup -C$ consists of the union of the two polyhedral cones C and $-C$. Note that $C \cup -C$ can be viewed as the set of all lines determined by the nonzero vectors in C (and $-C$). It is a double cone. Unless C is a closed half-space, $C \cup -C$ is not convex. See Figure 12.8. It seems perfectly natural to define a projective polyhedron as any set of lines induced by a set of the form $C \cup -C$, where C is a polyhedral cone.

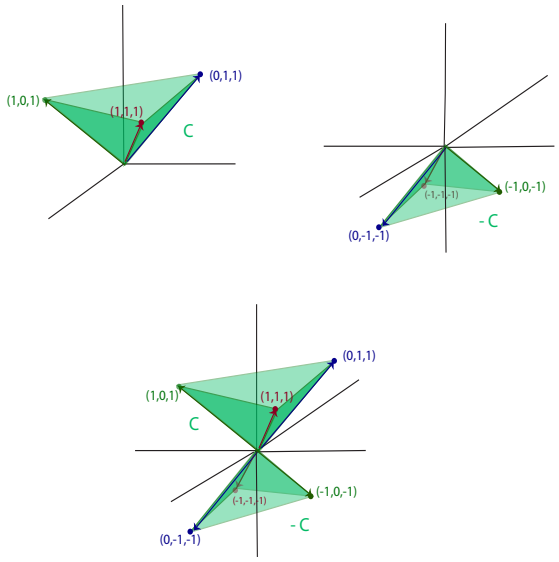


Figure 12.8: The double cone $C \cup -C$, where C is the $\mathcal{V}$-cone $C = \text{cone}\{(1,0,1), (0,1,1), (1,1,1)\}$.

Definition 12.9. A *projective polyhedron* is any subset, $P \subseteq \mathbb{P}^d$, of the form

$$P = p((C \cup -C) - \{0\}) = p(C - \{0\}),$$

where C is any polyhedral cone ($\mathcal{V}$ or $\mathcal{H}$ cone) in $\mathbb{R}^{d+1}$ (with $C \neq \{0\}$). We write $P = \mathbb{P}(C \cup -C)$ or $P = \mathbb{P}(C)$. See Figure 12.9.

It is important to observe that because $C \cup -C$ is a double cone there is a bijection between nontrivial double polyhedral cones and projective polyhedra. So, projective polyhedra are equivalent to double polyhedral cones. However, the projective interpretation of the lines induced by $C \cup -C$ as points in $\mathbb{P}^d$ makes the study of projective polyhedra geometrically more interesting.

Projective polyhedra inherit many of the properties of cones but we have to be careful because we are really dealing with double cones, $C \cup -C$, and not cones. As a consequence, there are a few unpleasant surprises, for example, the fact that the collection of projective polyhedra is **not** closed under intersection!

Before dealing with these issues, let us show that every "standard" polyhedron $P \subseteq \mathbb{E}^d$ has a natural projective completion, $\widetilde{P} \subseteq \mathbb{P}^d$, such that on the affine patch U_{d+1} (where $x_{d+1} \neq 0$), $\widetilde{P} \upharpoonright U_{d+1} = P$. For this, we use our theorem on the Polyhedron–Cone Correspondence (Theorem 5.20, part (2)).

Let $A = X + U$, where X is a set of points in $\mathbb{E}^d$ and U is a cone in $\mathbb{R}^d$. For every point,

$x \in X$, and every vector, $u \in U$, let

$$\widehat{x} = \begin{pmatrix} x \\ 1 \end{pmatrix}, \qquad \widehat{u} = \begin{pmatrix} u \\ 0 \end{pmatrix},$$

and let $\widehat{X} = \{\widehat{x} \mid x \in X\}$, $\widehat{U} = \{\widehat{u} \mid u \in U\}$ and $\widehat{A} = \{\widehat{a} \mid a \in A\}$, with

$$\widehat{a} = \begin{pmatrix} a \\ 1 \end{pmatrix}.$$

Then,

$$C(A) = \mathrm{cone}(\{\widehat{X} \cup \widehat{U}\})$$

is a cone in $\mathbb{R}^{d+1}$ such that

$$\widehat{A} = C(A) \cap H_{d+1},$$

where H_{d+1} is the hyperplane of equation $x_{d+1} = 1$. If we set $\widetilde{A} = \mathbb{P}(C(A))$, then we get a subset of $\mathbb{P}^d$ and in the patch U_{d+1}, the set $\widetilde{A} \restriction U_{d+1}$ is in bijection with the intersection $(C(A) \cup -C(A)) \cap H_{d+1} = \widehat{A}$, and thus, in bijection with A.

We call $\widetilde{A}$ the *projective completion of* A. We have an injection, $A \longrightarrow \widetilde{A}$, given by

$$(a_1, \ldots, a_d) \mapsto (a_1 \colon \cdots \colon a_d \colon 1),$$

which is just the map, $\psi_{d+1} \colon \mathbb{R}^d \to U_{d+1}$.

What the projective completion does is to add to A the "points at infinity" corresponding to the vectors in U, that is, the points of $\mathbb{P}^d$ corresponding to the lines in the cone, U.

Definition 12.10. If $X = \mathrm{conv}(Y)$ and $U = \mathrm{cone}(V)$ for some finite sets $Y = \{y_1, \ldots, y_p\}$ and $V = \{v_1, \ldots, v_q\}$, then $P = \mathrm{conv}(Y) + \mathrm{cone}(V)$ is a $\mathcal{V}$-polyhedron and $\widetilde{P} = \mathbb{P}(C(P))$ is a projective polyhedron. The projective polyhedron $\widetilde{P} = \mathbb{P}(C(P))$ is called the *projective completion* of P. See Figure 12.9.

Observe that if C is a closed half-space in $\mathbb{R}^{d+1}$, then $P = \mathbb{P}(C \cup -C) = \mathbb{P}^d$. Now, if $C \subseteq \mathbb{R}^{d+1}$ is a polyhedral cone and C is contained in a closed half-space, it is still possible that C contains some nontrivial linear subspace and we would like to understand this situation.

The first thing to observe is that $U = C \cap (-C)$ is the largest linear subspace contained in C.

Definition 12.11. If $C \cap (-C) = \{0\}$, we say that C is a *pointed* or *strongly convex* cone.

In this case, one immediately realizes that 0 is an extreme point of C and so, there is a hyperplane, H, through 0 so that $C \cap H = \{0\}$, that is, except for its apex, C lies in one of the open half-spaces determined by H. As a consequence, by a linear change of coordinates, we may assume that this hyperplane is $H_{d+1}(0)$ and so, for every projective polyhedron,

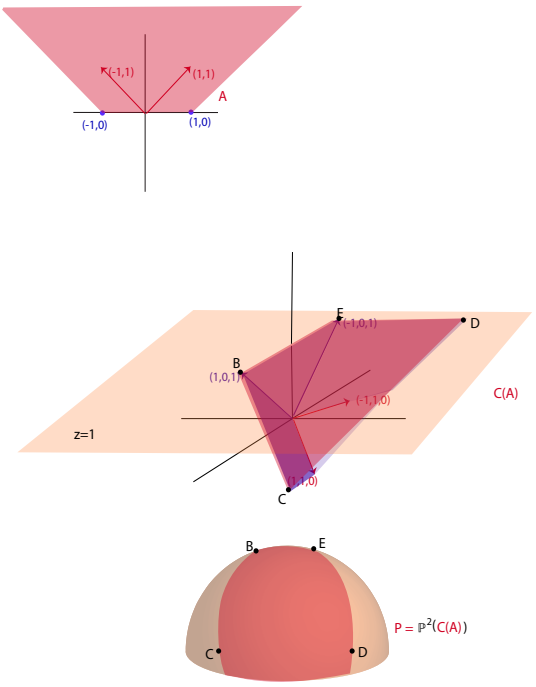


Figure 12.9: The bottom figure shows a projective polyhedron, which is the projective completion (in the halfsphere model of $\mathbb{P}^2$) of the infinite trough $A = X + U$, where $X = \{(-1,0),(1,0)\}$ and $U = \{(-1,1),(1,1)\}$.

$P = \mathbb{P}(C)$, if C is pointed then there is an affine patch (say, U_{d+1}) where P has no points at infinity, that is, P is a polytope! On the other hand, from another patch, U_i, as $P \upharpoonright U_i$ is in bijection with $(C \cup -C) \cap H_i$, the projective polyhedron P viewed on U_i may consist of *two* disjoint polyhedra.

The situation is very similar to the classical theory of projective conics or quadrics (for example, see Brannan, Esplen and Gray, [15]). The case where C is a pointed cone corresponds to the nondegenerate conics or quadrics. In the case of the conics, depending how we slice a cone, we see an ellipse, a parabola or a hyperbola.

For projective polyhedra, when we slice a polyhedral double cone, $C \cup -C$, we may see a polytope (*elliptic type*) a single unbounded polyhedron (*parabolic type*) or two unbounded polyhedra (*hyperbolic type*). See Figure 12.10.

Now, when $U = C \cap (-C) \neq \{0\}$, the polyhedral cone, C, contains the linear subspace, U, and if $C \neq \mathbb{R}^{d+1}$, then for every hyperplane, H, such that C is contained in one of the two closed half-spaces determined by H, the subspace $U \cap H$ is nontrivial. An example is the cone, $C \subseteq \mathbb{R}^3$, determined by the intersection of two planes through 0 (a wedge). In this case, U is equal to the line of intersection of these two planes. Also observe that $C \cap (-C) = C$

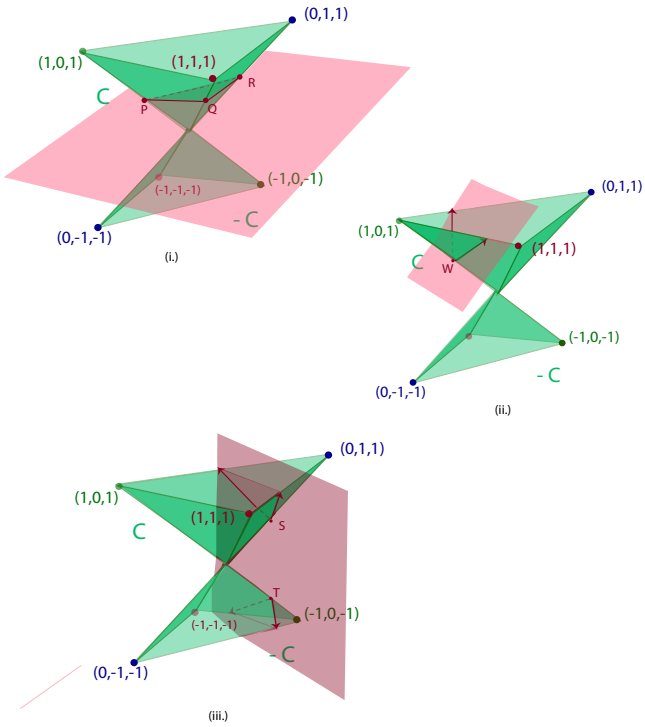


Figure 12.10: For the sea green double cone $C \cup -C$ of Figure 12.8, Figure (i.) illustrates an elliptic type polytope, Figure (ii.) illustrates a parabolic type polyhedron, while Figure (iii.) illustrates a hyperbolic type polyhedron.

iff $C = -C$, that is, iff C is a linear subspace.

The situation where $C \cap (-C) \neq \{0\}$ is reminiscent of the case of cylinders in the theory of quadric surfaces (see Brannan, Esplen and Gray [15] or Berger [8]). Now, every cylinder can be viewed as the ruled surface defined as the family of lines orthogonal to a plane and touching some nondegenerate conic.

A similar decomposition holds for polyhedral cones as shown below in a proposition borrowed from Ewald [26] (Chapter V, Lemma 1.6). We should warn the reader that we have some doubts about the proof given there, and so we offer a different proof adapted from the proof of Lemma 16.2 in Barvinok [4]. See Figure 12.11. Given any two subsets, $V, W \subseteq \mathbb{R}^d$, as usual, we write $V+W = \{v+w \mid v \in V, w \in W\}$ and $v+W = \{v+w \mid w \in W\}$, for any $v \in \mathbb{R}^d$.

Proposition 12.2. *For every polyhedral cone $C \subseteq \mathbb{R}^d$, if $U = C \cap (-C)$, then there is some pointed cone C_0 so that U and C_0 are orthogonal and*

$$C = U + C_0,$$

with $\dim(U) + \dim(C_0) = \dim(C)$.

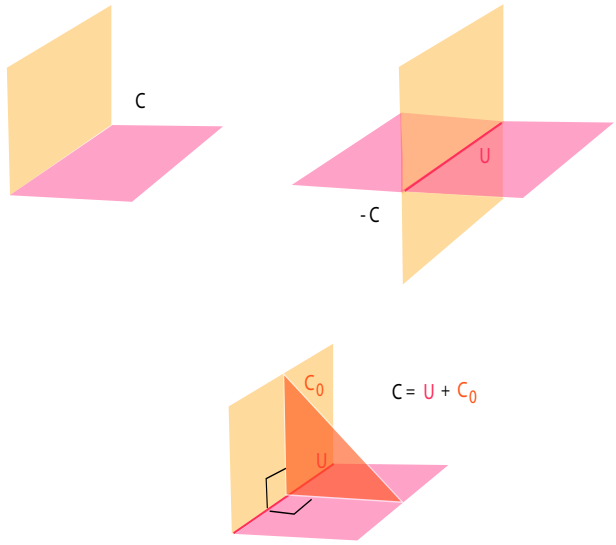


Figure 12.11: In $\mathbb{R}^3$, C is the cone determined by the pink and peach half planes, and $U = C \cap -C$ is the red line of intersection. Then $C = U + C_0$, where C_0 it the peach pointed cone contained in plane perpendicular to U.

Proof. We already know that $U = C \cap (-C)$ is the largest linear subspace of C. Let $U^\perp$ be the orthogonal complement of U in $\mathbb{R}^d$ and let $\pi\colon \mathbb{R}^d \to U^\perp$ be the orthogonal projection onto $U^\perp$. By Proposition 5.13, the projection, $C_0 = \pi(C)$, of C onto $U^\perp$ is a polyhedral cone. We claim that C_0 is pointed and that

$$C = U + C_0.$$

Since $\pi^{-1}(v) = v + U$ for every $v \in C_0$, we have $U + C_0 \subseteq C$. On the other hand, by definition of C_0, we also have $C \subseteq U + C_0$, so $C = U + C_0$. If C_0 was not pointed, then it would contain a linear subspace, V, of dimension at least 1 but then, $U + V$ would be a linear subspace of C of dimension strictly greater than U, which is impossible. Finally, $\dim(U) + \dim(C_0) = \dim(C)$ is obvious by orthogonality. □

Definition 12.12. The linear subspace $U = C \cap (-C)$ is called the *cospan* of C.

Both U and C_0 are uniquely determined by C. To a great extent, Proposition 12.2 reduces the study of non-pointed cones to the study of pointed cones.

Definition 12.13. We call the projective polyhedra of the form $P = \mathbb{P}(C)$, where C is a cone with a non-trivial cospan (a non-pointed cone) a *projective polyhedral cylinder*, by analogy with the quadric surfaces. We also propose to call the projective polyhedra of the form $P = \mathbb{P}(C)$, where C is a pointed cone, a *projective polytope* (or *nondegenerate projective polyhedron*).

The following propositions show that projective polyhedra behave well under projective maps and intersection with a hyperplane:

Proposition 12.3. *Given any projective map, $h\colon \mathbb{P}^m \to \mathbb{P}^n$, for any projective polyhedron, $P \subseteq \mathbb{P}^m$, the image, $h(P)$, of P is a projective polyhedron in $\mathbb{P}^n$. Even if $h\colon \mathbb{P}^m \to \mathbb{P}^n$ is a partial map but h is defined on P, then $h(P)$ is a projective polyhedron.*

Proof. The projective map, $h\colon \mathbb{P}^m \to \mathbb{P}^n$, is of the form $h = \mathbb{P}(\widehat{h})$, for some injective linear map, $\widehat{h}\colon \mathbb{R}^{m+1} \to \mathbb{R}^{n+1}$. Moreover, the projective polyhedron, P, is of the form $P = \mathbb{P}(C)$, for some polyhedral cone, $C \subseteq \mathbb{R}^{m+1}$, with $C = \text{cone}(\{u_1, \ldots, u_p\})$, for some nonzero vector $u_i \in \mathbb{R}^{m+1}$. By definition,

$$\mathbb{P}(h)(P) = \mathbb{P}(h)(\mathbb{P}(C)) = \mathbb{P}(\widehat{h}(C)).$$

As $\widehat{h}$ is linear,

$$\widehat{h}(C) = \widehat{h}(\text{cone}(\{u_1, \ldots, u_p\})) = \text{cone}(\{\widehat{h}(u_1), \ldots, \widehat{h}(u_p)\}).$$

If we let $\widehat{C} = \text{cone}(\{\widehat{h}(u_1), \ldots, \widehat{h}(u_p)\})$, then $\widehat{h}(C) = \widehat{C}$ is a polyhedral cone and so,

$$\mathbb{P}(h)(P) = \mathbb{P}(\widehat{h}(C)) = \mathbb{P}(\widehat{C})$$

is a projective cone. This argument does not depend on the injectivity of $\widehat{h}$, as long as $C \cap \text{Ker}\,(\widehat{h}) = \{0\}$. □

Proposition 12.3 together with earlier arguments shows that every projective polytope, $P \subseteq \mathbb{P}^d$, is equivalent under some suitable projectivity to another projective polytope, P', which is a polytope when viewed in the affine patch, U_{d+1}. This property is similar to the fact that every (non-degenerate) projective conic is projectively equivalent to an ellipse.

Since the notion of a face is defined for arbitrary polyhedra it is also defined for cones. Consequently, we can define the notion of a face for projective polyhedra.

Definition 12.14. Given a projective polyhedron $P \subseteq \mathbb{P}^d$, where $P = \mathbb{P}(C)$ for some polyhedral cone (uniquely determined by P) $C \subseteq \mathbb{R}^{d+1}$, a *face of P* is any subset of P of the form $\mathbb{P}(F) = p(F - \{0\})$, for any nontrivial face $F \subseteq C$ of C ($F \neq \{0\}$). Consequently, we say that $\mathbb{P}(F)$ is a *vertex* iff $\dim(F) = 1$, an *edge* iff $\dim(F) = 2$ and a *facet* iff $\dim(F) = \dim(C) - 1$. The projective polyhedron P and the empty set are the *improper* faces of P.

If C is strongly convex, then it is easy to prove that C is generated by its edges (its one-dimensional faces, these are rays) in the sense that any set of nonzero vectors spanning these edges generates C (using positive linear combinations). As a consequence, if C is strongly convex, we may say that P is "spanned" by its vertices, since P is equal to $\mathbb{P}$(all positive combinations of vectors representing its edges).

Remark: Even though we did not define the notion of convex combination of points in $\mathbb{P}^d$, the notion of projective polyhedron gives us a way to mimic certain properties of convex sets in the framework of projective geometry. That's because every projective polyhedron corresponds to a unique polyhedral cone.

If our projective polyhedron is the completion $\widetilde{P} = \mathbb{P}(C(P)) \subseteq \mathbb{P}^d$ of some polyhedron $P \subseteq \mathbb{R}^d$, then each face of the cone $C(P)$ is of the form $C(F)$, where F is a face of P and so, each face of $\widetilde{P}$ is of the form $\mathbb{P}(C(F))$, for some face F of P. In particular, in the affine patch U_{d+1} the face $\mathbb{P}(C(F))$ is in bijection with the face F of P. We will usually identify $\mathbb{P}(C(F))$ and F.

We now consider the intersection of projective polyhedra but first, let us make some general remarks about the intersection of subsets of $\mathbb{P}^d$. Given any two nonempty subsets, $\mathbb{P}(S)$ and $\mathbb{P}(S')$, of $\mathbb{P}^d$ where S and S' are polyhedral cones (or more generally cones with vertex 0), what is $\mathbb{P}(S) \cap \mathbb{P}(S')$? It is tempting to say that

$$\mathbb{P}(S) \cap \mathbb{P}(S') = \mathbb{P}(S \cap S'),$$

but unfortunately this is generally false! The problem is that $\mathbb{P}(S) \cap \mathbb{P}(S')$ is the set of *all lines* determined by vectors both in S and S' but there may be some line spanned by some vector $u \in (-S) \cap S'$ or $u \in S \cap (-S')$ such that u does not belong to $S \cap S'$ or $-(S \cap S')$.

Observe that

$$\begin{aligned} -(-S) &= S \\ -(S \cap S') &= (-S) \cap (-S'). \end{aligned}$$

Then, the correct intersection is given by

$$\begin{aligned} (S \cup -S) \cap (S' \cup -S') &= (S \cap S') \cup ((-S) \cap (-S')) \cup (S \cap (-S')) \cup ((-S) \cap S') \\ &= (S \cap S') \cup -(S \cap S') \cup (S \cap (-S')) \cup -(S \cap (-S')), \end{aligned}$$

which is the union of two double cones (except for 0, which belongs to both). Therefore, if $\mathbb{P}(S) \cap \mathbb{P}(S') \neq \emptyset$, then $S \cap S' \neq \{0\}$ or $S \cap (-S') \neq \{0\}$, and so

$$\mathbb{P}(S) \cap \mathbb{P}(S') = \mathbb{P}(S \cap S') \cup \mathbb{P}(S \cap (-S')) = \mathbb{P}(S \cap S') \cup \mathbb{P}((-S) \cap S'),$$

since $\mathbb{P}(S \cap (-S')) = \mathbb{P}((-S) \cap S')$, with the understanding that if $S \cap S = \{0\}$ or $S \cap (-S') = \{0\}$, then the corresponding term should be omitted.

Furthermore, if S' is symmetric (*i.e.*, $S' = -S'$), then

$$\begin{aligned} (S \cup -S) \cap (S' \cup -S') &= (S \cup -S) \cap S' \\ &= (S \cap S') \cup ((-S) \cap S') \\ &= (S \cap S') \cup -(S \cap (-S')) \\ &= (S \cap S') \cup -(S \cap S'). \end{aligned}$$

Thus, if either S or S' is symmetric and if $\mathbb{P}(S) \cap \mathbb{P}(S') \neq \emptyset$ then

$$\mathbb{P}(S) \cap \mathbb{P}(S') = \mathbb{P}(S \cap S').$$

Now, if C is a pointed polyhedral cone then $C \cap (-C) = \{0\}$. Consequently, for any other polyhedral cone C' we have $(C \cap C') \cap ((-C) \cap C') = \{0\}$. Using these facts and adopting the convention that $\mathbb{P}(\{0\}) = \emptyset$, we obtain the following result:

Proposition 12.4. *Let $P = \mathbb{P}(C)$ and $P' = \mathbb{P}(C')$ be any two projective polyhedra in $\mathbb{P}^d$. If $\mathbb{P}(C) \cap \mathbb{P}(C') \neq \emptyset$, then the following properties hold:*

(1)

$$\mathbb{P}(C) \cap \mathbb{P}(C') = \mathbb{P}(C \cap C') \cup \mathbb{P}(C \cap (-C')),$$

the union of two projective polyhedra. If C or C' is a pointed cone i.e., P or P' is a projective polytope, then $\mathbb{P}(C \cap C')$ and $\mathbb{P}(C \cap (-C'))$ are disjoint (if both are defined). See Figures 12.12 and 12.13.

(2) If $P' = H$ for some hyperplane $H \subseteq \mathbb{P}^d$, then $P \cap H$ is a projective polyhedron.

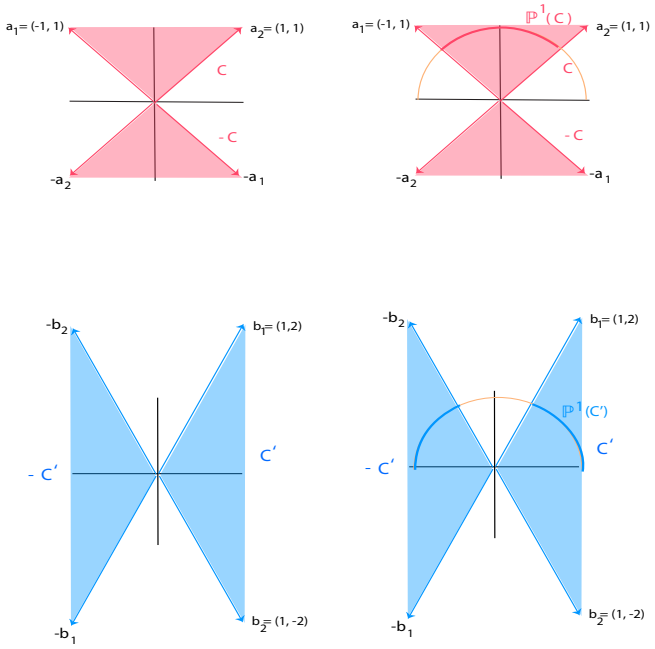


Figure 12.12: Let $C = \text{cone}\{(1,1),(-1,1)\}$ and $C' = \text{cone}\{(1,2),(1,-2)\}$. In the half-spherical model of $\mathbb{P}^1$, $\mathbb{P}(C)$ is the bold red arc, while $\mathbb{P}(C')$ is the bold blue arc.

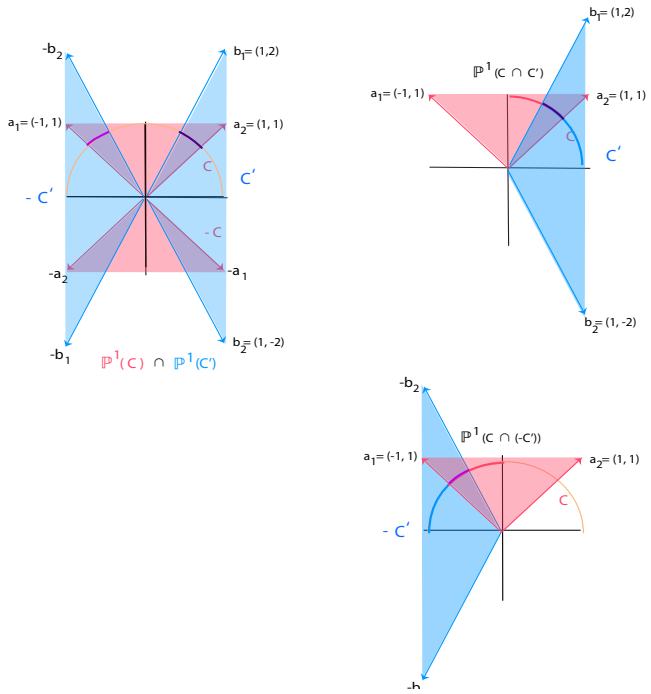


Figure 12.13: For the cones C and C' defined in Figure 12.12, $\mathbb{P}(C) \cap \mathbb{P}(C')$ is illustrated by two disjoint purple arcs; the light purple arc is $\mathbb{P}(C \cap C')$, while the dark purple arc is $\mathbb{P}(C \cap (-C'))$.

Proof. We already proved (1) so only (2) remains to be proved. Of course, we may assume that $P \neq \mathbb{P}^d$. This time, using the equivalence theorem of $\mathcal{V}$-cones and $\mathcal{H}$-cones (Theorem 5.19), we know that P is of the form $P = \mathbb{P}(C)$, with $C = \bigcap_{i=1}^{p} C_i$, where the C_i are closed half-spaces in $\mathbb{R}^{d+1}$. Moreover, $H = \mathbb{P}(\widehat{H})$, for some hyperplane, $\widehat{H} \subseteq \mathbb{R}^{d+1}$, through 0. Now, as $\widehat{H}$ is symmetric,

$$P \cap H = \mathbb{P}(C) \cap \mathbb{P}(\widehat{H}) = \mathbb{P}(C \cap \widehat{H}).$$

Consequently,

$$\begin{aligned} P \cap H &= \mathbb{P}(C \cap \widehat{H}) \\ &= \mathbb{P}\left(\left(\bigcap_{i=1}^{p} C_i\right) \cap \widehat{H}\right). \end{aligned}$$

However, $\widehat{H} = \widehat{H}_+ \cap \widehat{H}_-$, where $\widehat{H}_+$ and $\widehat{H}_-$ are the two closed half-spaces determined by $\widehat{H}$ and so,

$$\widehat{C} = \left(\bigcap_{i=1}^{p} C_i\right) \cap \widehat{H} = \left(\bigcap_{i=1}^{p} C_i\right) \cap \widehat{H}_+ \cap \widehat{H}_-$$

is a polyhedral cone. Therefore, $P \cap H = \mathbb{P}(\widehat{C})$ is a projective polyhedron. □

Proposition 12.4 can be sharpened a little.

Proposition 12.5. *Let $P = \mathbb{P}(C)$ and $P' = \mathbb{P}(C')$ be any two projective polyhedra in $\mathbb{P}^d$. If $\mathbb{P}(C) \cap \mathbb{P}(C') \neq \emptyset$, then*

$$\mathbb{P}(C) \cap \mathbb{P}(C') = \mathbb{P}(C \cap C') \cup \mathbb{P}(C \cap (-C')),$$

the union of two projective polyhedra. If $C = -C$, i.e., C is a linear subspace (or if C' is a linear subspace), then

$$\mathbb{P}(C) \cap \mathbb{P}(C') = \mathbb{P}(C \cap C').$$

Furthermore, if either C or C' is pointed, the above projective polyhedra are disjoint, else if C and C' both have nontrivial cospan and $\mathbb{P}(C \cap C')$ and $\mathbb{P}(C \cap (-C'))$ intersect then

$$\mathbb{P}(C \cap C') \cap \mathbb{P}(C \cap (-C')) = \mathbb{P}(C \cap (C' \cap (-C'))) \cup \mathbb{P}(C' \cap (C \cap (-C))).$$

Finally, if the two projective polyhedra on the right-hand side intersect, then

$$\mathbb{P}(C \cap (C' \cap (-C'))) \cap \mathbb{P}(C' \cap (C \cap (-C))) = \mathbb{P}((C \cap (-C)) \cap (C' \cap (-C'))).$$

Proof. Left as a simple exercise in boolean algebra. □

In preparation for Section 13.7, we also need the notion of tangent space at a point of a variety.

12.3 Tangent Spaces of Hypersurfaces and Projective Hypersurfaces

Since we only need to consider the case of hypersurfaces we restrict attention to this case (but the general case is a straightforward generalization). Let us begin with a hypersurface of equation $p(x_1, \ldots, x_d) = 0$ in $\mathbb{R}^d$, that is, the set

$$S = V(p) = \{(x_1, \ldots, x_d) \in \mathbb{R}^d \mid p(x_1, \ldots, x_d) = 0\},$$

where $p(x_1, \ldots, x_d)$ is a polynomial of total degree m.

Pick any point $a = (a_1, \ldots, a_d) \in \mathbb{R}^d$. Recall that there is a version of the Taylor expansion formula for polynomials such that, for any polynomial $p(x_1, \ldots, x_d)$ of total degree m, for every $h = (h_1, \ldots, h_d) \in \mathbb{R}^d$, we have

$$\begin{aligned} p(a+h) &= p(a) + \sum_{1 \leq |\alpha| \leq m} \frac{D^\alpha p(a)}{\alpha!} h^\alpha \\ &= p(a) + \sum_{i=1}^{d} p_{x_i}(a) h_i + \sum_{2 \leq |\alpha| \leq m} \frac{D^\alpha p(a)}{\alpha!} h^\alpha, \end{aligned}$$

where we use the *multi-index notation*, with $\alpha = (i_1, \ldots, i_d) \in \mathbb{N}^d$, $|\alpha| = i_1 + \cdots + i_d$, $\alpha! = i_1! \cdots i_d!$, $h^\alpha = h_1^{i_1} \cdots h_d^{i_d}$,

$$D^\alpha p(a) = \frac{\partial^{i_1+\cdots+i_d} p}{\partial x_1^{i_1} \cdots \partial x_d^{i_d}}\,(a),$$

and

$$p_{x_i}(a) = \frac{\partial p}{\partial x_i}\,(a).$$

Consider any line ℓ through a, given parametrically by

$$\ell = \{a + th \mid t \in \mathbb{R}\},$$

with $h \neq 0$ and say $a \in S$ is a point on the hypersurface $S = V(p)$, which means that $p(a) = 0$. The intuitive idea behind the notion of the tangent space to S at a is that it is the set of lines that intersect S at a in a point of *multiplicity at least two*, which means that the equation giving the intersection, $S \cap \ell$, namely

$$p(a + th) = p(a_1 + th_1, \ldots, a_d + th_d) = 0,$$

is of the form

$$t^2 q(a, h)(t) = 0,$$

where $q(a, h)(t)$ is some polynomial in t. Using Taylor's formula, as $p(a) = 0$, we have

$$p(a + th) = t \sum_{i=1}^{d} p_{x_i}(a) h_i + t^2 q(a, h)(t),$$

for some polynomial $q(a, h)(t)$. From this, we see that a is an intersection point of multiplicity at least 2 iff

$$\sum_{i=1}^{d} p_{x_i}(a) h_i = 0. \tag{$\dagger$}$$

Consequently, if $\nabla p(a) = (p_{x_1}(a), \ldots, p_{x_d}(a)) \neq 0$ (that is, if the gradient of p at a is nonzero), we see that ℓ intersects S at a in a point of multiplicity at least 2 iff h belongs to the hyperplane of equation ($\dagger$).

Definition 12.15. Let $S = V(p)$ be a hypersurface in $\mathbb{R}^d$. For any point $a \in S$, if $\nabla p(a) \neq 0$, then we say that a is a *non-singular* point of S. When a is nonsingular, the (*affine*) *tangent space* $T_a(S)$ (or simply $T_a S$) to S at a is the hyperplane through a of equation

$$\sum_{i=1}^{d} p_{x_i}(a)(x_i - a_i) = 0.$$

See Figure 12.14.

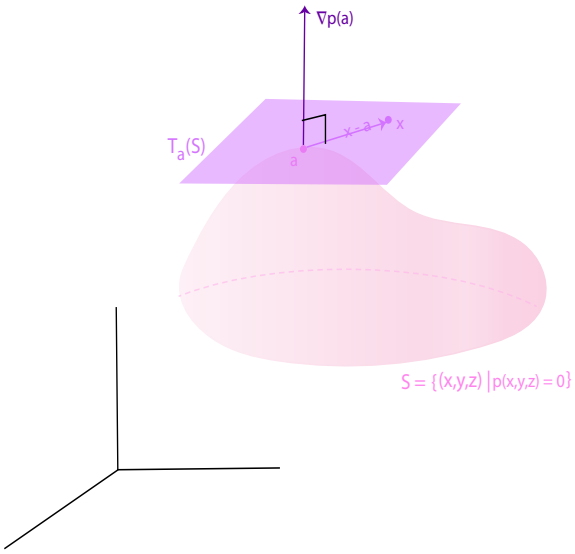


Figure 12.14: The tangent plane to the surface $S = V(p)$.

Observe that the hyperplane of the direction of T_aS is the hyperplane through 0 and parallel to T_aS given by

$$\sum_{i=1}^{d} p_{x_i}(a)x_i = 0.$$

When $\nabla p(a) = 0$, we either say that T_aS is undefined or we set $T_aS = \mathbb{R}^d$.

We now extend the notion of tangent space to projective varieties. As we will see, this amounts to homogenizing and the result turns out to be simpler than the affine case!

Definition 12.16. Let $S = V(F) \subseteq \mathbb{P}^d$ be a projective hypersurface, which means that

$$S = V(F) = \{(x_1 : \cdots : x_{d+1}) \in \mathbb{P}^d \mid F(x_1, \ldots, x_{d+1}) = 0\},$$

where $F(x_1, \ldots, x_{d+1})$ is a homogeneous polynomial of total degree m. We say that a point $a \in S$ is *non-singular* iff $\nabla F(a) = (F_{x_1}(a), \ldots, F_{x_{d+1}}(a)) \neq 0$.

For every $i = 1, \ldots, d+1$, let

$$z_j^{\restriction i} = \frac{x_j}{x_i},$$

where $j = 1, \ldots, d+1$ and $j \neq i$, and let $f^{\restriction i}$ be the result of "dehomogenizing" F at i, that is,

$$f^{\restriction i}(z_1^{\restriction i}, \ldots, z_{i-1}^{\restriction i}, z_{i+1}^{\restriction i}, \ldots, z_{d+1}^{\restriction i}) = F(z_1^{\restriction i}, \ldots, z_{i-1}^{\restriction i}, 1, z_{i+1}^{\restriction i}, \ldots, z_{d+1}^{\restriction i}).$$

Definition 12.17. We define the *(projective) tangent space* T_aS to a at S as the hyperplane H such that for each affine patch U_i where $a_i \neq 0$, if we let

$$a_j^{\restriction i} = \frac{a_j}{a_i},$$

where $j = 1, \ldots, d+1$ and $j \neq i$, then the restriction $H \restriction U_i$ of H to U_i is the affine hyperplane tangent to $S \restriction U_i$ given by

$$\sum_{\substack{j=1 \\ j\neq i}}^{d+1} f^{\restriction i}_{z_j^{\restriction i}}(a^{\restriction i})(z_j^{\restriction i} - a_j^{\restriction i}) = 0.$$

Thus, on the affine patch U_i, the tangent space T_aS is given by the homogeneous equation

$$\sum_{\substack{j=1 \\ j\neq i}}^{d+1} f^{\restriction i}_{z_j^{\restriction i}}(a^{\restriction i})(x_j - a_j^{\restriction i} x_i) = 0.$$

This looks awful but we can make it pretty if we remember that F is a homogeneous polynomial of degree m and that we have the *Euler relation*:

$$\sum_{j=1}^{d+1} F_{x_j}(a) a_j = mF(a),$$

for every $a = (a_1, \ldots, a_{d+1}) \in \mathbb{R}^{d+1}$. Using this, we can come up with a clean equation for our projective tangent hyperplane. It is enough to carry out the computations for $i = d+1$.

Our tangent hyperplane has the equation

$$\sum_{j=1}^{d} F_{x_j}(a_1^{\restriction d+1}, \ldots, a_d^{\restriction d+1}, 1)(x_j - a_j^{\restriction d+1} x_{d+1}) = 0,$$

that is,

$$\sum_{j=1}^{d} F_{x_j}(a_1^{\restriction d+1}, \ldots, a_d^{\restriction d+1}, 1) x_j + \sum_{j=1}^{d} F_{x_j}(a_1^{\restriction d+1}, \ldots, a_d^{\restriction d+1}, 1)(-a_j^{\restriction d+1} x_{d+1}) = 0.$$

As $F(x_1, \ldots, x_{d+1})$ is homogeneous of degree m, and as $a_{d+1} \neq 0$ on U_{d+1}, we have

$$a_{d+1}^m F(a_1^{\restriction d+1}, \ldots, a_d^{\restriction d+1}, 1) = F(a_1, \ldots, a_d, a_{d+1}),$$

so from the above equation we get

$$\sum_{j=1}^{d} F_{x_j}(a_1, \ldots, a_{d+1}) x_j + \sum_{j=1}^{d} F_{x_j}(a_1, \ldots, a_{d+1})(-a_j^{\restriction d+1} x_{d+1}) = 0. \qquad (*)$$

Since $a \in S$, we have $F(a) = 0$, so the Euler relation yields

$$\sum_{j=1}^{d} F_{x_j}(a_1, \ldots, a_{d+1})a_j + F_{x_{d+1}}(a_1, \ldots, a_{d+1})a_{d+1} = 0,$$

which, by dividing by a_{d+1} and multiplying by x_{d+1}, yields

$$\sum_{j=1}^{d} F_{x_j}(a_1, \ldots, a_{d+1})(-a_j^{\restriction d+1} x_{d+1}) = F_{x_{d+1}}(a_1, \ldots, a_{d+1})x_{d+1},$$

and by plugging this in $(*)$, we get

$$\sum_{j=1}^{d} F_{x_j}(a_1, \ldots, a_{d+1})x_j + F_{x_{d+1}}(a_1, \ldots, a_{d+1})x_{d+1} = 0.$$

Consequently, the tangent hyperplane to S at a is given by the equation

$$\sum_{j=1}^{d+1} F_{x_j}(a)x_j = 0.$$

Definition 12.18. Let $S = V(F)$ be a hypersurface in $\mathbb{P}^d$, where $F(x_1, \ldots, x_{d+1})$ is a homogeneous polynomial. For any point $a \in S$, if $\nabla F(a) \neq 0$, then we say that a is a *non-singular* point of S. When a is nonsingular, the (*projective*) *tangent space* $T_a(S)$ (or simply $T_a S$) to S at a is the hyperplane through a of equation

$$\sum_{i=1}^{d+1} F_{x_i}(a)x_i = 0.$$

For example, if we consider the sphere $S^2 \subseteq \mathbb{P}^3$ of equation

$$x^2 + y^2 + z^2 - w^2 = 0,$$

the tangent plane to S^2 at $a = (a_1, a_2, a_3, a_4)$ is given by

$$a_1 x + a_2 y + a_3 z - a_4 w = 0.$$

Remark: If $a \in S = V(F)$, as $F(a) = \sum_{i=1}^{d+1} F_{x_i}(a)a_i = 0$ (by Euler), the equation of the tangent plane $T_a S$ to S at a can also be written as

$$\sum_{i=1}^{d+1} F_{x_i}(a)(x_i - a_i) = 0.$$

Now, if $a = (a_1 \colon \cdots \colon a_d \colon 1)$ is a point in the affine patch U_{d+1}, then the equation of the intersection of T_aS with U_{d+1} is obtained by setting $a_{d+1} = x_{d+1} = 1$, that is

$$\sum_{i=1}^{d} F_{x_i}(a_1, \ldots, a_d, 1)(x_i - a_i) = 0,$$

which is just the equation of the affine hyperplane to $S \cap U_{d+1}$ at $a \in U_{d+1}$.

It will be convenient to adopt the following notational convention: Given any point $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ written as a row vector, we let $\mathbf{x}$ denote the corresponding column vector such that $\mathbf{x}^\top = x$.

Projectivities behave well with respect to hypersurfaces and their tangent spaces. Let $S = V(F) \subseteq \mathbb{P}^d$ be a projective hypersurface, where F is a homogeneous polynomial of degree m and let $h \colon \mathbb{P}^d \to \mathbb{P}^d$ be a projectivity (a bijective projective map). Assume that h is induced by the invertible $(d+1) \times (d+1)$ matrix, $A = (a_{i\,j})$, and write $A^{-1} = (a^{-1}_{i\,j})$. For any hyperplane, $H \subseteq \mathbb{R}^{d+1}$, if φ is any linear from defining φ, *i.e.*, $H = \operatorname{Ker}(\varphi)$, then

$$\begin{aligned} h(H) &= \{h(x) \in \mathbb{R}^{d+1} \mid \varphi(x) = 0\} \\ &= \{y \in \mathbb{R}^{d+1} \mid (\exists x \in \mathbb{R}^{d+1})(y = h(x),\ \varphi(x) = 0)\} \\ &= \{y \in \mathbb{R}^{d+1} \mid (\varphi \circ h^{-1})(y) = 0\}. \end{aligned}$$

Consequently, if H is given by

$$\alpha_1 x_1 + \cdots + \alpha_{d+1} x_{d+1} = 0$$

and if we write $\alpha = (\alpha_1, \ldots, \alpha_{d+1})$, then $h(H)$ is the hyperplane given by the equation

$$\alpha A^{-1} \mathbf{y} = 0.$$

Similarly,

$$\begin{aligned} h(S) &= \{h(x) \in \mathbb{R}^{d+1} \mid F(x) = 0\} \\ &= \{y \in \mathbb{R}^{d+1} \mid (\exists x \in \mathbb{R}^{d+1})(y = h(x),\ F(x) = 0)\} \\ &= \{y \in \mathbb{R}^{d+1} \mid F((A^{-1}\mathbf{y})^\top) = 0\} \end{aligned}$$

is the hypersurface defined by the polynomial

$$G(x_1, \ldots, x_{d+1}) = F\left(\sum_{j=1}^{d+1} a^{-1}_{1\,j} x_j, \ldots, \sum_{j=1}^{d+1} a^{-1}_{d+1\,j} x_j\right).$$

Furthermore, using the chain rule, we get

$$(G_{x_1}, \ldots, G_{x_{d+1}}) = (F_{x_1}, \ldots, F_{x_{d+1}})A^{-1},$$

which shows that a point, $a \in S$, is non-singular iff its image, $h(a) \in h(S)$, is non-singular on $h(S)$. This also shows that

$$h(T_aS) = T_{h(a)}h(S),$$

that is, the projectivity, h, preserves tangent spaces. In summary, we have

Proposition 12.6. *Let $S = V(F) \subseteq \mathbb{P}^d$ be a projective hypersurface, where F is a homogeneous polynomial of degree m, and let $h\colon \mathbb{P}^d \to \mathbb{P}^d$ be a projectivity (a bijective projective map). Then, $h(S)$ is a hypersurface in $\mathbb{P}^d$ and a point $a \in S$ is nonsingular for S iff $h(a)$ is nonsingular for $h(S)$. Furthermore,*

$$h(T_a S) = T_{h(a)} h(S),$$

that is, the projectivity h preserves tangent spaces.

Remark: If $h\colon \mathbb{P}^m \to \mathbb{P}^n$ is a projective map, say induced by an injective linear map given by the $(n+1) \times (m+1)$ matrix $A = (a_{i\,j})$, given any hypersurface $S = V(F) \subseteq \mathbb{P}^n$, we can define the *pull-back* $h^*(S) \subseteq \mathbb{P}^m$ of S, by

$$h^*(S) = \{x \in \mathbb{P}^m \mid F(h(x)) = 0\}.$$

This is indeed a hypersurface because $F(x_1, \ldots, x_{n+1})$ is a homogenous polynomial and $h^*(S)$ is the zero locus of the homogeneous polynomial

$$G(x_1, \ldots, x_{m+1}) = F\left(\sum_{j=1}^{m+1} a_{1\,j} x_j, \ldots, \sum_{j=1}^{m+1} a_{n+1\,j} x_j\right).$$

If $m = n$ and h is a projectivity, then we have

$$h(S) = (h^{-1})^*(S).$$

12.4 Quadrics (Affine, Projective) and Polar Duality

The case where $S = V(\Phi) \subseteq \mathbb{P}^d$ is a hypersurface given by a homogeneous polynomial $\Phi(x_1, \ldots, x_{d+1})$ of degree 2 will come up a lot and deserves a little more attention. In this case, if we write $x = (x_1, \ldots, x_{d+1})$, then $\Phi(x) = \Phi(x_1, \ldots, x_{d+1})$ is completely determined by a $(d+1) \times (d+1)$ symmetric matrix, say $F = (f_{i\,j})$, and we have

$$\Phi(x) = \mathbf{x}^\top F \mathbf{x} = \sum_{i,j=1}^{d+1} f_{i\,j} x_i x_j.$$

Since F is symmetric, we can write

$$\Phi(x) = \sum_{i,j=1}^{d+1} f_{i\,j} x_i x_j = \sum_{i=1}^{d+1} f_{i\,i} x_i^2 + 2 \sum_{\substack{i,j=1 \\ i<j}}^{d+1} f_{i\,j} x_i x_j.$$

Definition 12.19. The *polar form* $\varphi(x, y)$ of $\Phi(x)$, is given by

$$\varphi(x, y) = \mathbf{x}^\top F \mathbf{y} = \sum_{i,j=1}^{d+1} f_{ij} x_i y_j,$$

where $x = (x_1, \ldots, x_{d+1})$ and $y = (y_1, \ldots, y_{d+1})$.

Of course,

$$2\varphi(x, y) = \Phi(x + y) - \Phi(x) - \Phi(y).$$

We also check immediately that

$$2\varphi(x, y) = 2\mathbf{x}^\top F \mathbf{y} = \sum_{j=1}^{d+1} \frac{\partial \Phi(x)}{\partial x_j}\, y_j,$$

and so,

$$\left(\frac{\partial \Phi(x)}{\partial x_1}, \ldots, \frac{\partial \Phi(x)}{\partial x_{d+1}}\right) = 2\mathbf{x}^\top F.$$

Definition 12.20. The hypersurface $S = V(\Phi) \subseteq \mathbb{P}^d$ is called a *(projective) (hyper-)quadric surface*. We say that a quadric surface $S = V(\Phi)$ is *nondegenerate* iff the matrix F defining Φ is invertible.

For example, the sphere, $S^d \subseteq \mathbb{P}^{d+1}$, is the nondegenerate quadric given by

$$\mathbf{x}^\top \begin{pmatrix} I_{d+1} & \mathbf{0} \\ \mathbb{O} & -1 \end{pmatrix} \mathbf{x} = 0$$

and the paraboloid, $\mathcal{P} \subseteq \mathbb{P}^{d+1}$, is the nongenerate quadric given by

$$\mathbf{x}^\top \begin{pmatrix} I_d & \mathbf{0} & \mathbf{0} \\ \mathbb{O} & 0 & -\frac{1}{2} \\ \mathbb{O} & -\frac{1}{2} & 0 \end{pmatrix} \mathbf{x} = 0.$$

If $h\colon \mathbb{P}^d \to \mathbb{P}^d$ is a projectivity induced by some invertible matrix, $A = (a_{ij})$, and if $S = V(\Phi)$ is a quadric defined by the matrix F, we immediately check that $h(S)$ is the quadric defined by the matrix $(A^{-1})^\top F A^{-1}$. Furthermore, as A is invertible, we see that S is nondegenerate iff $h(S)$ is nondegenerate.

Observe that polar duality w.r.t. the sphere, S^{d-1}, can be expressed by

$$X^* = \left\{ x \in \mathbb{R}^d \mid (\forall y \in X) \left((\mathbf{x}^\top, 1) \begin{pmatrix} I_d & \mathbf{0} \\ \mathbb{O} & -1 \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ 1 \end{pmatrix} \leq 0 \right) \right\},$$

where X is any subset of $\mathbb{R}^d$. The above suggests generalizing polar duality with respect to any nondegenerate quadric.

Let $Q = V(\Phi(x_1, \ldots, x_{d+1}))$ be a nondegenerate quadric given by the homogeneous polynomial Φ with corresponding polar form φ and matrix $F = (f_{ij})$. Then, we know that φ induces a natural duality between $\mathbb{R}^{d+1}$ and $(\mathbb{R}^{d+1})^*$, namely, for every $u \in \mathbb{R}^{d+1}$, if $\varphi_u \in (\mathbb{R}^{d+1})^*$ is the linear form given by

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

for every $v \in \mathbb{R}^{d+1}$, then the map $u \mapsto \varphi_u$, from $\mathbb{R}^{d+1}$ to $(\mathbb{R}^{d+1})^*$, is a linear isomorphism.

Definition 12.21. Let $Q = V(\Phi(x_1, \ldots, x_{d+1}))$ be a nondegenerate quadric with corresponding polar form φ. For any $u \in \mathbb{R}^{d+1}$, with $u \neq 0$, the set

$$u^\dagger = \{v \in \mathbb{R}^{d+1} \mid \varphi(u, v) = 0\} = \{v \in \mathbb{R}^{d+1} \mid \varphi_u(v) = 0\} = \text{Ker } \varphi_u$$

is a hyperplane called the *polar of u (w.r.t. Q)*. In terms of the matrix representation of Q, the polar of u is given by the equation

$$\mathbf{u}^\top F \mathbf{x} = 0,$$

or

$$\sum_{j=1}^{d+1} \frac{\partial \Phi(u)}{\partial x_j} x_j = 0.$$

Going over to $\mathbb{P}^d$, we say that $\mathbb{P}(u^\dagger)$ is the *polar (hyperplane)* of the point $a = [u] \in \mathbb{P}^d$ and we write $a^\dagger$ for $\mathbb{P}(u^\dagger)$.

Note that the equation of the polar hyperplane $a^\dagger$ of a point $a \in \mathbb{P}^d$ is identical to the equation of the tangent plane to Q at a, except that a is not necessarily on Q. However, if $a \in Q$, then the polar of a is indeed the tangent hyperplane $T_a Q$ to Q at a.

Proposition 12.7. *Let $Q = V(\Phi(x_1, \ldots, x_{d+1})) \subseteq \mathbb{P}^d$ be a nondegenerate quadric with corresponding polar form, φ, and matrix, F. Then, every point, $a \in Q$, is nonsingular.*

Proof. Since

$$\left(\frac{\partial \Phi(a)}{\partial x_1}, \ldots, \frac{\partial \Phi(a)}{\partial x_{d+1}}\right) = 2\mathbf{a}^\top F,$$

if $a \in Q$ is singular, then $\mathbf{a}^\top F = 0$ with $a \neq 0$, contradicting the fact that F is invertible. □

The reader should prove the following simple proposition:

Proposition 12.8. *Let $Q = V(\Phi(x_1, \ldots, x_{d+1}))$ be a nondegenerate quadric with corresponding polar form, φ. Then, the following properties hold: For any two points, $a, b \in \mathbb{P}^d$,*

(1) $a \in b^\dagger$ iff $b \in a^\dagger$;

(2) $a \in a^\dagger$ iff $a \in Q$;

(3) Q does not contain any hyperplane.

Remark: As in the case of the sphere, if Q is a nondegenerate quadric and $a \in \mathbb{P}^d$ is any point such that the polar hyperplane $a^\dagger$ intersects Q, then there is a nice geometric interpretation for $a^\dagger$. Observe that for every $b \in Q \cap a^\dagger$, the polar hyperplane $b^\dagger$ is the tangent hyperplane T_bQ to Q at b and that $a \in T_bQ$. Also, if $a \in T_bQ$ for any $b \in Q$, as $b^\dagger = T_bQ$, then $b \in a^\dagger$. Therefore, $Q \cap a^\dagger$ is the set of contact points of all the tangent hyperplanes to Q passing through a.

Proposition 12.9. *Every hyperplane $H \subseteq \mathbb{P}^d$ is the polar of a single point $a \in \mathbb{P}^d$.*

Proof. Indeed, if H is defined by a nonzero linear form $f \in (\mathbb{R}^{d+1})^*$, as Φ is nondegenerate, there is a unique $u \in \mathbb{R}^{d+1}$, with $u \neq 0$, so that $f = \varphi_u$, and as φ_u vanishes on H, we see that H is the polar of the point $a = [u]$. If H is also the polar of another point $b = [v]$, then φ_v vanishes on H, which means that

$$\varphi_v = \lambda\varphi_u = \varphi_{\lambda u},$$

with $\lambda \neq 0$ and this implies $v = \lambda u$, that is, $a = [u] = [v] = b$, and the pole of H is indeed unique. □

Definition 12.22. Let $Q = V(\Phi(x_1, \ldots, x_{d+1}))$ be a nondegenerate quadric with corresponding polar form φ. The *polar dual (w.r.t. Q)* X^* of a subset $X \subseteq \mathbb{R}^{d+1}$ is given by

$$X^* = \{v \in \mathbb{R}^{d+1} \mid (\forall u \in X)(\varphi(u, v) \leq 0)\}.$$

For every subset $X \subseteq \mathbb{P}^d$, we let

$$X^* = \mathbb{P}((v(X))^*),$$

where $v(X)$ is the unique double cone associated with X as in Proposition 12.1.

Observe that X^* is always a cone, even if $X \subseteq \mathbb{R}^{d+1}$ is not. By analogy with the Euclidean case, for any nonzero vector $u \in \mathbb{R}^{d+1}$, let

$$(u^\dagger)_- = \{v \in \mathbb{R}^{d+1} \mid \varphi(u, v) \leq 0\}.$$

Now, we have the following version of Proposition 5.5:

Proposition 12.10. *Let $Q = V(\Phi(x_1, \ldots, x_{d+1}))$ be a nondegenerate quadric with corresponding polar form φ, and matrix $F = (f_{ij})$. For any nontrivial polyhedral cone $C = \mathrm{cone}(u_1, \ldots, u_p)$, where $u_i \in \mathbb{R}^{d+1}$, $u_i \neq 0$, we have*

$$C^* = \bigcap_{i=1}^{p} (u_i^\dagger)_-.$$

If U is the $(d+1) \times p$ matrix whose i^{th} column is u_i, then we can also write

$$C^* = P(U^\top F, \mathbf{0}),$$

where

$$P(U^\top F, \mathbf{0}) = \{v \in \mathbb{R}^{d+1} \mid U^\top F \mathbf{v} \leq \mathbf{0}\}.$$

Consequently, the polar dual of a polyhedral cone w.r.t. a nondegenerate quadric is a polyhedral cone.

Proof. The proof is essentially the same as the proof of Proposition 5.5. As

$$C = \text{cone}(u_1, \ldots, u_p) = \{\lambda_1 u_1 + \cdots + \lambda_p u_p \mid \lambda_i \geq 0,\ 1 \leq i \leq p\},$$

we have

$$\begin{aligned}
C^* &= \{v \in \mathbb{R}^{d+1} \mid (\forall u \in C)(\varphi(u, v) \leq 0)\} \\
&= \{v \in \mathbb{R}^{d+1} \mid \varphi(\lambda_1 u_1 + \cdots + \lambda_p u_p, v) \leq 0,\ \lambda_i \geq 0,\ 1 \leq i \leq p\} \\
&= \{v \in \mathbb{R}^{d+1} \mid \lambda_1 \varphi(u_1, v) + \cdots + \lambda_p \varphi(u_p, v) \leq 0,\ \lambda_i \geq 0,\ 1 \leq i \leq p\} \\
&= \bigcap_{i=1}^{p} \{v \in \mathbb{R}^{d+1} \mid \varphi(u_i, v) \leq 0\} \\
&= \bigcap_{i=1}^{p} (u_i^\dagger)_-.
\end{aligned}$$

By the equivalence theorem for $\mathcal{H}$-polyhedra and $\mathcal{V}$-polyhedra, we conclude that C^* is a polyhedral cone. □

Proposition 12.10 allows us to make the following definition:

Definition 12.23. Let $Q = V(\Phi(x_1, \ldots, x_{d+1}))$ be a nondegenerate quadric with corresponding polar form φ. Given any projective polyhedron $P = \mathbb{P}(C)$, where C is a polyhedral cone, the *polar dual (w.r.t. Q)* P^* of P is the projective polyhedron

$$P^* = \mathbb{P}(C^*).$$

We also show that projectivities behave well with respect to polar duality.

Proposition 12.11. *Let $Q = V(\Phi(x_1, \ldots, x_{d+1}))$ be a nondegenerate quadric with corresponding polar form φ, and matrix $F = (f_{i\,j})$. For every projectivity $h\colon \mathbb{P}^d \to \mathbb{P}^d$, if h is induced by the linear map $\widehat{h}$ given by the invertible matrix $A = (a_{i\,j})$, for every subset $X \subseteq \mathbb{R}^{d+1}$, we have*

$$\widehat{h}(X^*) = (\widehat{h}(X))^*,$$

where on the left-hand side X^* is the polar dual of X w.r.t. Q, and on the right-hand side $(\widehat{h}(X))^*$ is the polar dual of $\widehat{h}(X)$ w.r.t. the nondegenerate quadric $h(Q)$ given by the matrix $(A^{-1})^\top F A^{-1}$. Consequently, if $X \neq \{0\}$, then

$$h((\mathbb{P}(X))^*) = (h(\mathbb{P}(X)))^*$$

and for every projective polyhedron P, we have

$$h(P^*) = (h(P))^*.$$

Proof. As

$$X^* = \{v \in \mathbb{R}^{d+1} \mid (\forall u \in X)(\mathbf{u}^\top F \mathbf{v} \leq 0)\},$$

we have

$$\begin{aligned}
\widehat{h}(X^*) &= \{\widehat{h}(v) \in \mathbb{R}^{d+1} \mid (\forall u \in X)(\mathbf{u}^\top F \mathbf{v} \leq 0)\} \\
&= \{y \in \mathbb{R}^{d+1} \mid (\forall u \in X)(\mathbf{u}^\top F A^{-1} \mathbf{y} \leq 0)\} \\
&= \{y \in \mathbb{R}^{d+1} \mid (\forall x \in \widehat{h}(X))(\mathbf{x}^\top (A^{-1})^\top F A^{-1} \mathbf{y} \leq 0)\} \\
&= (\widehat{h}(X))^*,
\end{aligned}$$

where $(\widehat{h}(X))^*$ is the polar dual of $\widehat{h}(X)$ w.r.t. the quadric whose matrix is $(A^{-1})^\top F A^{-1}$, that is, the polar dual w.r.t. $h(Q)$.

The second part of the proposition follows immediately by setting $X = C$, where C is the polyhedral cone defining the projective polyhedron, $P = \mathbb{P}(C)$. □

We will also need the notion of an affine quadric and polar duality with respect to an affine quadric. Fortunately, the properties we need in the affine case are easily derived from the projective case using the "trick" that the affine space $\mathbb{E}^d$ can be viewed as the hyperplane $H_{d+1} \subseteq \mathbb{R}^{d+1}$ of equation, $x_{d+1} = 1$, and that its associated vector space $\mathbb{R}^d$ can be viewed as the hyperplane $H_{d+1}(0) \subseteq \mathbb{R}^{d+1}$ of equation $x_{d+1} = 0$. A point, $a \in \mathbb{A}^d$, corresponds to the vector $\widehat{a} = \binom{a}{1} \in \mathbb{R}^{d+1}$, and a vector $u \in \mathbb{R}^d$ corresponds to the vector, $\widehat{u} = \binom{u}{0} \in \mathbb{R}^{d+1}$. This way, the projective space $\mathbb{P}^d = \mathbb{P}(\mathbb{R}^{d+1})$ is the natural *projective completion* of $\mathbb{E}^d$, which is isomorphic to the affine patch U_{d+1} where $x_{d+1} \neq 0$. The hyperplane $x_{d+1} = 0$ is the "hyperplane at infinity" in $\mathbb{P}^d$.

If we write $x = (x_1, \ldots, x_d)$, a polynomial, $\Phi(x) = \Phi(x_1, \ldots, x_d)$, of degree 2 can be written as

$$\Phi(x) = \sum_{i,j=1}^{d} a_{ij} x_i x_j + 2 \sum_{i=1}^{d} b_i x_i + c,$$

where $A = (a_{ij})$ is a symmetric matrix. If we write $b^\top = (b_1, \ldots, b_d)$, then we have

$$\Phi(x) = (\mathbf{x}^\top, 1) \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} = \widehat{\mathbf{x}}^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} \widehat{\mathbf{x}}.$$

Therefore, as in the projective case, Φ is completely determined by a $(d+1)\times(d+1)$ symmetric matrix, say $F=(f_{i\,j})$, and we have

$$\Phi(x)=(\mathbf{x}^\top,1)F\begin{pmatrix}\mathbf{x}\\1\end{pmatrix}=\widehat{\mathbf{x}}^\top F\widehat{\mathbf{x}}.$$

Definition 12.24. We say that $Q\subseteq\mathbb{R}^d$ is a *nondegenerate affine quadric* iff

$$Q=V(\Phi)=\left\{x\in\mathbb{R}^d\mid(\mathbf{x}^\top,1)F\begin{pmatrix}\mathbf{x}\\1\end{pmatrix}=0\right\},$$

where F is symmetric and invertible. Given any point $a\in\mathbb{R}^d$, the *polar hyperplane* $a^\dagger$ of a w.r.t. Q is defined by

$$a^\dagger=\left\{x\in\mathbb{R}^d\mid(\mathbf{a}^\top,1)F\begin{pmatrix}\mathbf{x}\\1\end{pmatrix}=0\right\}.$$

From a previous discussion, the equation of the polar hyperplane $a^\dagger$ is

$$\sum_{i=1}^{d}\frac{\partial\Phi(a)}{\partial x_i}\,(x_i-a_i)=0.$$

Definition 12.25. Given any subset $X\subseteq\mathbb{R}^d$, the *polar dual* X^* of X is defined by

$$X^*=\left\{y\in\mathbb{R}^d\mid(\forall x\in X)\left((\mathbf{x}^\top,1)F\begin{pmatrix}\mathbf{y}\\1\end{pmatrix}\leq 0\right)\right\}.$$

As noted before, polar duality with respect to the affine sphere $S^d\subseteq\mathbb{R}^{d+1}$ corresponds to the case where

$$F=\begin{pmatrix}I_d&\mathbf{0}\\\mathbb{O}&-1\end{pmatrix},$$

and polar duality with respect to the affine paraboloid $\mathcal{P}\subseteq\mathbb{R}^{d+1}$ corresponds to the case where

$$F=\begin{pmatrix}I_{d-1}&\mathbf{0}&\mathbf{0}\\\mathbb{O}&0&-\frac{1}{2}\\\mathbb{O}&-\frac{1}{2}&0\end{pmatrix}.$$

We will need the following version of Proposition 5.15:

Proposition 12.12. *Let Q be a nondegenerate affine quadric given by the $(d+1)\times(d+1)$ symmetric matrix F, 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 $\widehat{Y}$ is the $(d+1)\times p$ matrix whose i^{th} column is $\widehat{y_i}$ and $\widehat{V}$ is the $(d+1)\times q$ matrix whose j^{th} column is $\widehat{v_j}$, then*

$$(\mathrm{conv}(\{y_1,\ldots,y_p\})\cup\mathrm{cone}(\{v_1,\ldots,v_q\}))^*=P(\widehat{Y}^\top F,\mathbf{0};\widehat{V}^\top F,\mathbf{0}),$$

with

$$P(\widehat{Y}^\top F,\mathbf{0};\widehat{V}^\top F,\mathbf{0})=\left\{x\in\mathbb{R}^d\mid\widehat{Y}^\top F\begin{pmatrix}\mathbf{x}\\1\end{pmatrix}\leq\mathbf{0},\ \widehat{V}^\top F\begin{pmatrix}\mathbf{x}\\0\end{pmatrix}\leq\mathbf{0}\right\}.$$

Proof. The proof is immediately adapted from that of Proposition 5.15. □

Using Proposition 12.12, we can prove the following Proposition showing that projective completion and polar duality commute:

Proposition 12.13. *Let Q be a nondegenerate affine quadric given by the $(d+1)\times(d+1)$ symmetric, invertible matrix F. For every polyhedron $P \subseteq \mathbb{R}^d$, we have*

$$\widetilde{P^*} = (\widetilde{P})^*,$$

where on the right-hand side, we use polar duality w.r.t. the nondegenerate projective quadric $\widetilde{Q}$ defined by F.

Proof. By definition, we have $\widetilde{P} = \mathbb{P}(C(P))$, $(\widetilde{P})^* = \mathbb{P}((C(P))^*)$ and $\widetilde{P^*} = \mathbb{P}(C(P^*))$. Therefore, it suffices to prove that

$$(C(P))^* = C(P^*).$$

Now, $P = \mathrm{conv}(Y) + \mathrm{cone}(V)$, for some finite set of points Y and some finite set of vectors V, and we know that

$$C(P) = \mathrm{cone}(\widehat{Y} \cup \widehat{V}).$$

From Proposition 12.10,

$$(C(P))^* = \{v \in \mathbb{R}^{d+1} \mid \widehat{Y}^\top F\mathbf{v} \leq \mathbf{0},\ \widehat{V}^\top F\mathbf{v} \leq \mathbf{0}\}$$

and by Proposition 12.12,

$$P^* = \left\{x \in \mathbb{R}^d \mid \widehat{Y}^\top F\begin{pmatrix}\mathbf{x}\\ 1\end{pmatrix} \leq \mathbf{0},\ \widehat{V}^\top F\begin{pmatrix}\mathbf{x}\\ 0\end{pmatrix} \leq \mathbf{0}\right\}.$$

But, by definition of $C(P^*)$ (see Section 5.5, especially Proposition 5.20), the hyperplanes cutting out $C(P^*)$ are obtained by homogenizing the equations of the hyperplanes cutting out P^* and so,

$$C(P^*) = \left\{\begin{pmatrix}\mathbf{x}\\ x_{d+1}\end{pmatrix} \in \mathbb{R}^{d+1} \mid \widehat{Y}^\top F\begin{pmatrix}\mathbf{x}\\ x_{d+1}\end{pmatrix} \leq \mathbf{0},\ \widehat{V}^\top F\begin{pmatrix}\mathbf{x}\\ x_{d+1}\end{pmatrix} \leq \mathbf{0}\right\} = (C(P))^*,$$

as claimed. □

Remark: If $Q = V(\Phi(x_1, \ldots, x_{d+1}))$ is a projective or an affine quadric, it is obvious that

$$V(\Phi(x_1, \ldots, x_{d+1})) = V(\lambda\Phi(x_1, \ldots, x_{d+1}))$$

for every $\lambda \neq 0$. This raises the following question: If

$$Q = V(\Phi_1(x_1, \ldots, x_{d+1})) = V(\Phi_2(x_1, \ldots, x_{d+1})),$$

what is the relationship between Φ_1 and Φ_2?

The answer depends crucially on the field over which projective space or affine space is defined (*i.e.*, whether $Q \subseteq \mathbb{RP}^d$ or $Q \subseteq \mathbb{CP}^d$ in the projective case or whether $Q \subseteq \mathbb{R}^{d+1}$ or $Q \subseteq \mathbb{C}^{d+1}$ in the affine case).

For example, over $\mathbb{R}$, the polynomials $\Phi_1(x_1, x_2, x_3) = x_1^2 + x_2^2$ and $\Phi_2(x_1, x_2, x_3) = 2x_1^2 + 3x_2^2$ both define the point $(0\colon 0\colon 1) \in \mathbb{P}^2$, since the only real solution of Φ_1 and Φ_2 are of the form $(0, 0, z)$. However, if Q has some nonsingular point, the following can be proved (see Samuel [52], Theorem 46 (Chapter 3)):

Theorem 12.14. *Let $Q = V(\Phi(x_1, \ldots, x_{d+1})$ be a projective or an affine quadric over $\mathbb{RP}^d$ or $\mathbb{R}^{d+1}$. If Q has a nonsingular point, then for every polynomial Φ' such that $Q = V(\Phi'(x_1, \ldots, x_{d+1})$, there is some $\lambda \neq 0$ ($\lambda \in \mathbb{R}$) so that $\Phi' = \lambda\Phi$.*

In particular, Theorem 12.14 shows that the equation of a nondegenerate quadric is unique up to a scalar.

Actually, more is true. It turns out that if we allow complex solutions, that is, if $Q \subseteq \mathbb{CP}^d$ in the projective case or $Q \subseteq \mathbb{C}^{d+1}$ in the affine case, then $Q = V(\Phi_1) = V(\Phi_2)$ always implies $\Phi_2 = \lambda\Phi_1$ for some $\lambda \in \mathbb{C}$, with $\lambda \neq 0$. In the real case, the above holds (for some $\lambda \in \mathbb{R}$, with $\lambda \neq 0$) unless Q is an affine subspace (resp. a projective subspace) of dimension at most $d - 1$ (resp. of dimension at most $d - 2$). Even in this case, there is a bijective affine map f (resp. a bijective projective map h) such that $\Phi_2 = \Phi_1 \circ f^{-1}$ (resp. $\Phi_2 = \Phi_1 \circ h^{-1}$). A proof of these facts (and more) can be found in Tisseron [64] (Chapter 3).

We now have everything we need for a rigorous presentation of the material of Section 13.7. For a comprehensive treatment of the affine and projective quadrics and related material, the reader should consult Berger (Geometry II) [8] or Samuel [52].

Chapter 13

Dirichlet–Voronoi Diagrams and Delaunay Triangulations

In this chapter we present 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 of Voronoi diagrams and Delaunay triangulations is far from thorough. We are primarily interested in defining these concepts and stating their most important properties. For a comprehensive exposition of Voronoi diagrams, Delaunay triangulations, and more topics in computational geometry, our readers may consult O'Rourke [46], Preparata and Shamos [49], Boissonnat and Yvinec [12], de Berg, Van Kreveld, Overmars, and Schwarzkopf [6], or Risler [50]. The survey by Graham and Yao [34] contains a very gentle and lucid introduction to computational geometry.

In Section 13.7 (which relies on Sections 13.5 and 13.6), we show that the Delaunay triangulation of a set of points P is the stereographic projection of the convex hull of the set of points obtained by mapping the points in P onto the sphere using inverse stereographic projection. We also prove in Section 13.8 that the Voronoi diagram of P is obtained by taking the polar dual of the above convex hull and projecting it from the north pole (back onto the hyperplane containing P). A rigorous proof of this second fact is not trivial because the central projection from the north pole is only a partial map. To give a rigorous proof, we have to use projective completions. This requires defining convex polyhedra in projective space, and we use the results of Chapter 12 (especially, Section 12.2).

13.1 Dirichlet–Voronoi Diagrams

Let $\mathcal{E}$ be a Euclidean space of finite dimension, that is, an affine space $\mathcal{E}$ whose underlying vector space $\overrightarrow{\mathcal{E}}$ is equipped with an inner product (and has finite dimension). For concrete-

ness, one may safely assume that $\mathcal{E} = \mathbb{E}^m$, 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 13.1.

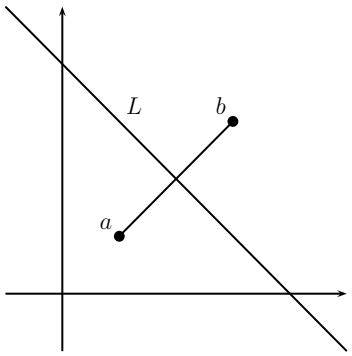


Figure 13.1: The bisector line L of a and b

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 **hm** is orthogonal to **ab**. In any orthogonal frame, letting $m = (x, y)$, $a = (a_1, a_2)$, $b = (b_1, b_2)$, the equation of this line is

$$(b_1 - a_1)(x - (a_1 + b_1)/2) + (b_2 - a_2)(y - (a_2 + b_2)/2) = 0,$$

which can also 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 [34]. 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 is

$$(b_1 - a_1)(x - (a_1 + b_1)/2) + (b_2 - a_2)(y - (a_2 + b_2)/2) + (b_3 - a_3)(z - (a_3 + b_3)/2) = 0,$$

which can also 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 [23] (1850) and Voronoi [68] (1908). Voronoi diagrams also arise in crystallography (Gilbert [32]). Other applications, including facility location and path planning, are discussed in O'Rourke [46]. For simplicity, we also denote the set $V(p_i)$ by V_i, and we introduce the following definition.

Definition 13.1. Let $\mathcal{E}$ be a Euclidean space of dimension $m = 2, 3$. Given any set $P = \{p_1, \ldots, p_n\}$ of n points in $\mathcal{E}$, the *Dirichlet–Voronoi diagram* $\mathcal{V}or(P)$ 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 $\mathbf{hm}$ is orthogonal to $\mathbf{ab}$. The equation of this hyperplane is

$$(b_1 - a_1)(x_1 - (a_1 + b_1)/2) + \cdots + (b_m - a_m)(x_m - (a_m + b_m)/2) = 0,$$

which can also 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

$$(b_1 - a_1)x_1 + \cdots + (b_m - a_m)x_m \geq (b_1^2 + \cdots + b_m^2)/2 - (a_1^2 + \cdots + a_m^2)/2.$$

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 13.2 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 $\mathcal{V}or(P)$ of P is the same as Definition 13.1, except that $H(p_i, p_j)$ is the closed half-space containing p_i and having the bisector hyperplane of p_i and p_j as boundary. Also, observe that the convex hull of P is a convex polytope.

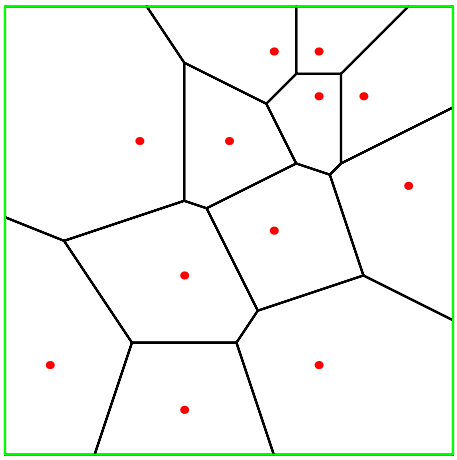


Figure 13.2: A Voronoi diagram

We will now state a proposition listing the main properties of Voronoi diagrams. It turns out that certain degenerate situations can be avoided if we assume that the points in the set P are in general position.

Definition 13.2. If P is a set of points in an affine space of dimension m, then we say that the points of P are in *general position* if no $m + 2$ points from P belong to the same $(m - 1)$-sphere.

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.

Proposition 13.1. *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)

$$\bigcup_{i=1}^{n} V_i = \mathcal{E}, \quad \textit{and} \quad \overset{\circ}{V}_i \cap \overset{\circ}{V}_j = \emptyset, \quad \textit{for all } i, j, \textit{ with } i \neq j,$$

where $\overset{\circ}{V}_i$ denotes the interior of V_i.

Proof. We prove only some of the statements, leaving the others as an exercise (or see Risler [50]).

(1) Since $V_i = \bigcap_{j \neq i} H(p_i, p_j)$ and each half-space $H(p_i, p_j)$ is convex, as an intersection of convex sets, V_i is convex. Also, since p_i belongs to the interior of each $H(p_i, p_j)$, the point p_i belongs to the interior of V_i.

(2) Let $F_{i,j}$ denote $V_i \cap V_j$. Any vertex p of the Vononoi diagram of P must belong to r faces $F_{i,j}$. Let us pick the origin of our affine space to be p. Now, given a vector space E and any two subspaces M and N of E, recall that we have the *Grassmann relation*

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

Then since p belongs to the intersection of hyperplanes that support the boundaries of the V_i, and since a hyperplane has dimension $m-1$, by the Grassmann relation, in order to obtain $\{p\}$, a subspace of dimension 0, as the intersection of hyperplanes, we must intersect at least m hyperplanes, so we must have $r \geq m$. We can rename the $r+1$ points p_i corresponding the regions V_i inducing the faces containing p by $p_1, \ldots, p_{r+1}$, so that the r faces containing p are denoted $F_{1,2}, F_{2,3}, \ldots, F_{r,r+1}$. Since $F_{i,j} = V_i \cap V_j$, we have

$$F_{i,j} = \{p \mid d(p, p_i) = d(p, p_j) \leq d(p, p_k), \text{ for all } k \neq i, j\},$$

and since $p \in F_{1,2} \cap F_{2,3} \cap \cdots \cap F_{r,r+1}$, we have

$$d(p, p_1) = \cdots = d(p, p_{r+1}) < d(p, p_k) \text{ for all } k \notin \{1, \ldots, r+1\}.$$

This means that p is the center of a sphere passing through $p_1, \ldots, p_{r+1}$ and containing no other point in P. By the assumption that points in P are in general position, since there are $r+1$ points p_i on a sphere, we must have $r+1 \leq m+1$, that is, $r \leq m$, and thus $r = m$. Thus, p belongs to $V_1 \cap \cdots \cap V_{m+1}$, but to no other V_j with $j \notin \{1, \ldots, m+1\}$. Furthermore, every edge of the Voronoi diagram containing p is the intersection of m of the regions $V_1, \ldots, V_{m+1}$, and so there are $m+1$ of them. □

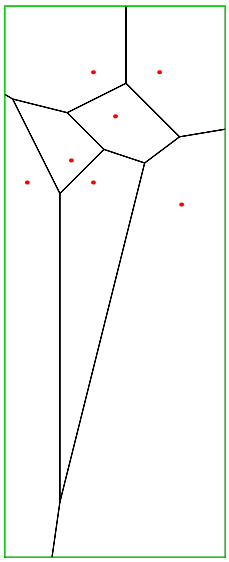


Figure 13.3: Another Voronoi diagram

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 13.3 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 $H(p_i, p_j)$ (with p_i fixed). One way to do this is to construct for each p_i successive convex polygons that converge to the boundary of the region $V(p_i)$. 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 [49] or O'Rourke [46]), 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 13.4, 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 $\mathcal{V}or(P)$), 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 13.4 shows the Delaunay triangulation associated with the earlier Voronoi diagram of a set of twelve points.

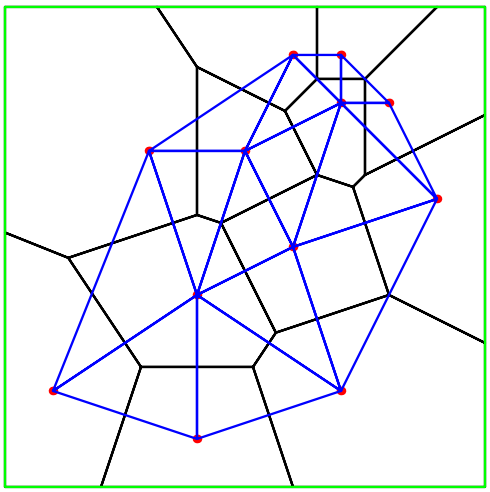


Figure 13.4: Delaunay triangulation associated with a Voronoi diagram

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. In Figure 13.5 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 any dimension $m \geq 3$.

13.2 Triangulations

The concept of a triangulation relies on the notion of pure simplicial complex defined in Chapter 10. The reader should review Definition 10.2 and Definition 10.3.

Definition 13.3. 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

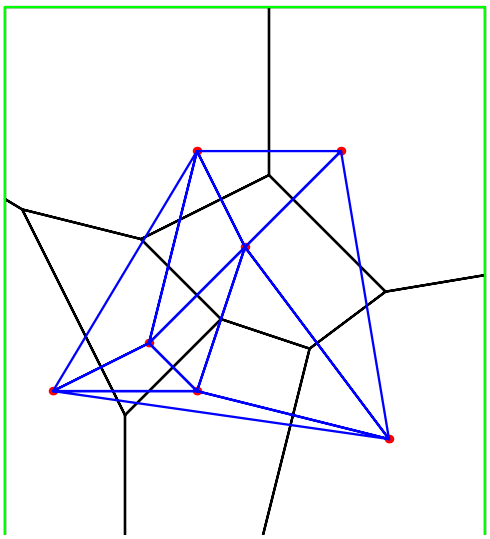


Figure 13.5: Another Delaunay triangulation associated with a Voronoi diagram

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).

Proposition 13.2.

(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 11.6. 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. Indeed, in the case of a planar triangulation, each face has 3 edges, and if there are e_b edges in the boundary and e_i edges not in the boundary, each nonboundary edge is shared by two faces, and thus $3f = e_b + 2e_i$. Since $v - e_b - e_i + f = 1$, 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.

13.3 Delaunay Triangulations

Given a set $P = \{p_1, \ldots, p_n\}$ of n points in the plane and the Voronoi diagram $\mathcal{V}or(P)$ for P, we explained in Section 13.1 how to define an (undirected) graph: The vertices of this graph are the points p_i (each corresponding to a unique region of $\mathcal{V}or(P)$), 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 13.4. Let $P = \{p_1, \ldots, p_n\}$ be a set of n points in $\mathbb{E}^m$, and let $\mathcal{V}or(P)$ be the Voronoi diagram of P. We define a complex $\mathcal{D}el(P)$ as follows. The complex $\mathcal{D}el(P)$ 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 $\mathcal{D}el(P)$ 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 $\mathcal{D}el$ instead of $\mathcal{D}el(P)$. A Delaunay triangulation for a set of twelve points is shown in Figure 13.6.

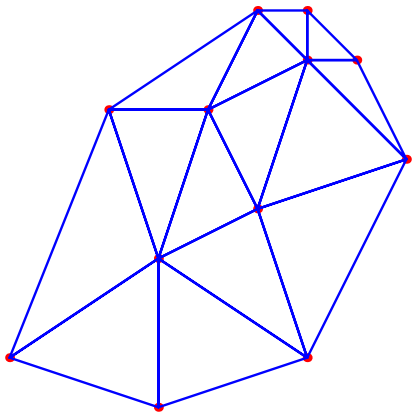


Figure 13.6: A Delaunay triangulation

Actually, it is not obvious that $\mathcal{D}el(P)$ is a triangulation of the convex hull of P, but this can be shown, as well as the properties listed in the following proposition.

Proposition 13.3. *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 $\mathcal{D}el(P)$ is the convex hull of P.

(2) A triangulation T associated with P is the Delaunay triangulation $\mathcal{D}el(P)$ 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 [50] and O'Rourke [46]. 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 [46]).

We will now explore briefly the connection between Delaunay triangulations and convex hulls.

13.4 Delaunay Triangulations and Convex Hulls

In this section we show that there is an intimate relationship between convex hulls and Delaunay triangulations. 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 $\mathbb{E}^{m+1}$ (a hypersurface), 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 Edelsbrunner and Seidel [24]. 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). Indeed, if such a circle C is defined by the equation

$$x^2 + y^2 + ax + by + c = 0,$$

since $X = x$, $Y = y$, and $Z = x^2 + y^2$, by eliminating $x^2 + y^2$ we get

$$Z = -ax - by - c,$$

and thus X, Y, Z satisfy the linear equation

$$aX + bY + Z + c = 0,$$

which is the equation of a plane. Thus, 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) as illustrated in Figure 13.7.

We can compute the convex hull of the set of lifted points. Let us focus on the downward-facing faces of this convex hull.

A downward-facing face is a face such that the z-coordinate of the unit normal to the plane supporting this face pointing towards the interior of the convex hull of the set of lifted points is positive.

Let $(l(p_1), l(p_2), l(p_3))$ be a downward-facing face, where the points p_1, p_2, p_3 belong to the set P. We claim that no other point from P is inside the circle C circumscribed about p_1, p_2, p_3.

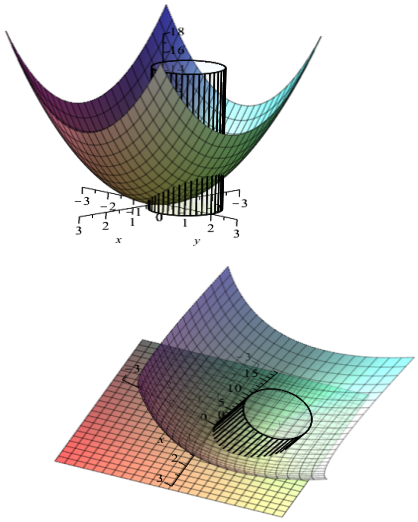


Figure 13.7: The intersection of the paraboloid $x^2+y^2 = z$ with the cylinder $x^2+(y-1)^2 = 0$. The intersection is an ellipse in the plane $z = 2y - 1$.

Indeed, a point p inside the circle C would lift to a point $l(p)$ on the paraboloid. Since no four points are cocyclic, one of the four points p_1, p_2, p_3, p is further from O than the others; say this point is p_3. Then, the face $(l(p_1), l(p_2), l(p))$ would be below the face $(l(p_1), l(p_2), l(p_3))$, contradicting the fact that $(l(p_1), l(p_2), l(p_3))$ is one of the downward-facing faces of the convex hull of P. See Figure 13.8. But then, by Property (2) of Proposition 13.3, the triangle (p_1, p_2, p_3) would belong to the Delaunay triangulation of P.

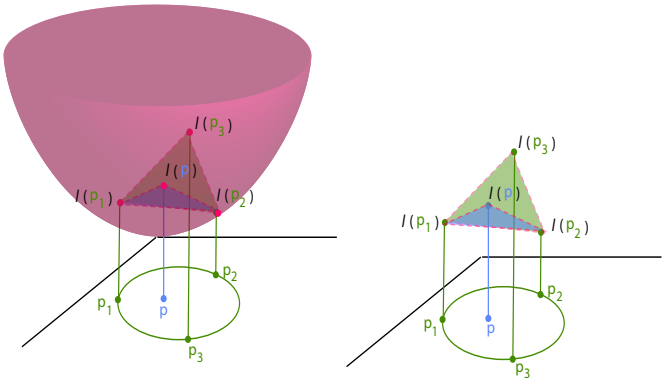


Figure 13.8: The lift of four points p_1, p_2, p_3, p. Since p is inside the green circle, the blue triangle $(l(p_1), l(p), l(p_2))$ is beneath the green triangle $(l(p_1), l(p_2), l(p_3))$, which implies that $(l(p_1), l(p_2), l(p_3))$ is not downward facing.

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

13.9 shows the lifting of the Delaunay triangulation shown earlier. Another example of the lifting of a Delaunay triangulation is shown in Figure 13.10.

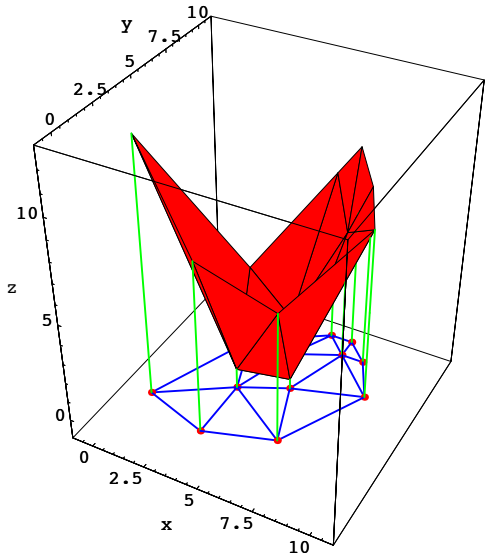


Figure 13.9: A Delaunay triangulation and its lifting to a paraboloid

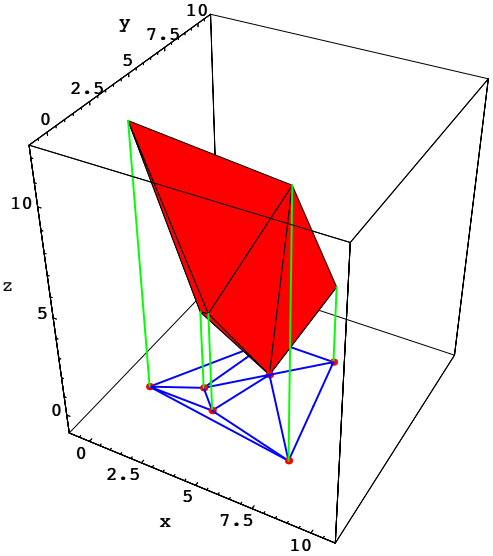


Figure 13.10: Another Delaunay triangulation and its lifting to a paraboloid

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.

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 the above is precisely the equation of the bisector line of the two points (a_1, b_1) and (a_2, b_2). See Figure 13.11. Therefore, *if we look at the paraboloid from $z = +\infty$ (with the paraboloid transparent), the projection of the boundary of the polyhedron $\mathcal{V}(P)$ consisting of the intersection of the half spaces containing the origin cut out by the tangent planes at the lifted points is the Voronoi diagram*!

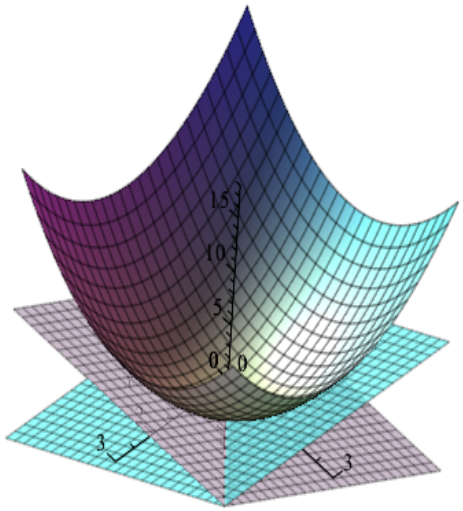


Figure 13.11: The intersection of the tangent plane at $(0, 1, 1)$ with equation $z = 2y - 1$, and the tangent plane at $(1, 0, 1)$ with equation $z = 2x - 1$, has intersection $y - x = 0$, namely the bisecting hyperplane between $(0, 1, 0)$ and $(1, 0, 0)$.

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 boundary of the polyhedron $\mathcal{V}(P)$ cut out by the tangent planes at the points of the lifted set $l(P)$ viewed from $z = +\infty$, is reminiscent of the polar duality with respect to a quadric. This duality will be thoroughly investigated in Section 13.7.

The reader interested in algorithms for finding Voronoi diagrams and Delaunay triangulations is referred to O'Rourke [46], Preparata and Shamos [49], Boissonnat and Yvinec [12], de Berg, Van Kreveld, Overmars, and Schwarzkopf [6], and Risler [50].

13.5 Stereographic Projection and the Space of Generalized Spheres

We saw in Section 13.4 that lifting a set of points $P \subseteq \mathbb{E}^2$ to the paraboloid $\mathcal{P}$ *via* the lifting function $l\colon \mathbb{E}^2 \to \mathcal{P}$ given by $l(x, y) = (x, y, x^2 + y^2)$ yields a definition of the Delaunay triangulation $\mathcal{D}el(P)$ of the set of points P that does not require any knowledge of the Voronoi diagram of P. Namely, $\mathcal{D}el(P)$ is the orthogonal projection of the part of the convex hull of the lifted set $l(P)$ consisting of its downward-facing faces. The Voronoi diagram $\mathcal{V}or(P)$ is also obtained from the lifted set $l(P)$; it is the projection of the boundary of the polyhedron $\mathcal{V}(P)$ cut out by the tangent planes at the points of the lifted set $l(P)$.

The generalization to any dimension $d \geq 2$ is immediate. Recall that the paraboloid $\mathcal{P}$ in $\mathbb{E}^{d+1}$ is given by the equation

$$x_{d+1} = \sum_{i=1}^{d} x_i^2,$$

and of course, the sphere S^d is given by

$$\sum_{i=1}^{d+1} x_i^2 = 1.$$

Then the lifting map $l\colon \mathbb{E}^d \to \mathcal{P}$ is given by

$$l(x_1, \ldots, x_d) = (x_1, \ldots, x_d, \sum_{i=1}^{d} x_i^2),$$

and the orthogonal projection $p_{d+1}\colon \mathbb{E}^{d+1} \to \mathbb{E}^d$ is given by

$$p_{d+1}(x_1, \ldots, x_d, x_{d+1}) = (x_1, \ldots, x_d).$$

As far as we know, Edelsbrunner and Seidel [24] were the first to find the relationship between Voronoi diagrams and the polar dual of the convex hull of a lifted set of points onto a paraboloid. This connection is described in Note 3.1 of Section 3 in [24]. The connection between the Delaunay triangulation and the convex hull of the lifted set of points is described in Note 3.2 of the same paper. Polar duality is not mentioned and seems to enter the scene only with Boissonnat and Yvinec [12].

Brown appears to be the first person who observed that Voronoi diagrams and convex hulls are related *via* inversion with respect to a sphere [16]. Brown takes a set of points P, for simplicity assumed to be in the plane, first lifts these points to the unit sphere S^2 using inverse stereographic projection from the north pole $\tau_N\colon \mathbb{E}^2 \to (S^2 - \{N\})$ (which is equivalent to an inversion of power 2 centered at the north pole), getting $\tau_N(P)$, and then takes the convex hull $\mathcal{D}(P) = \text{conv}(\tau_N(P))$ of the lifted set. Now, in order to obtain the

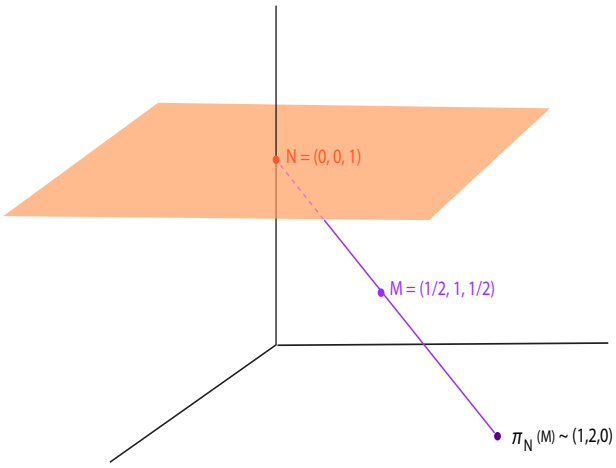


Figure 13.12: The geometric construction for $\pi_N \colon (\mathbb{E}^3 - H_3) \to \mathbb{E}^2$.

Voronoi diagram of P, apply our inversion (of power 2 centered at the north pole) to each of the faces of $\mathrm{conv}(\tau_N(P))$, obtaining spheres passing through the north pole, and then intersect these spheres with the plane containing P, obtaining circles. The centers of some of these circles are the Voronoi vertices. Finally, a simple criterion can be used to retain the "nearest Voronoi points" and to connect up these vertices; see Brown [16], page 225.

Note that Brown's method is *not* the method that uses the polar dual of the polyhedron $\mathcal{D}(P) = \mathrm{conv}(\tau_N(P))$, as we might have expected from the lifting method using a paraboloid. However, Brown's method suggests a method for obtaining the Delaunay triangulation $\mathcal{D}el(P)$ of P by lifting the set P to the sphere S^d by applying the inverse stereographic projection $\tau_N \colon \mathbb{E}^d \to (S^d - \{N\})$ (see Definition 13.5) instead of the lifting function l, computing the convex hull $\mathcal{D}(P) = \mathrm{conv}(\tau_N(P))$ of the lifted set $\tau_N(P)$, and then applying the central projection π_N from the north pole N to the hyperplane $x_{d+1} = 0$ instead of the orthogonal projection p_{d+1} to the facets of the polyhedron $\mathcal{D}(P)$ that do not contain the north pole, as we will prove in Section 13.7. The central projection π_N is the partial map $\pi_N \colon (\mathbb{E}^{d+1} - H_{d+1}) \to \mathbb{E}^d$ given by

$$\pi_N(x_1, \ldots, x_d, x_{d+1}) = \frac{1}{1 - x_{d+1}}(x_1, \ldots, x_d);$$

see Definition 13.5. For any point $M = (x_1, \ldots, x_d, x_{d+1})$ not in the hyperplane H_{d+1} of equation $x_{d+1} = 1$, the point $\pi_N(M)$ is the intersection of the line $\langle N, M \rangle$ through M and N with the hyperplane $H_{d+1}(0)$ of equation $x_{d+1} = 0$. See Figure 13.12.

Thus, instead of using a paraboloid we can use a sphere, and instead of the lifting function l we can use the the inverse stereographic projection τ_N. Then, to get back down to $\mathbb{E}^d$, we

use the central projection π_N instead of the orthogonal projection p_{d+1}. As $\mathcal{D}(P)$ is strictly below the hyperplane $x_{d+1} = 1$, there are no problems.

It turns out that there is a "projective transformation" Θ of $\mathbb{E}^{d+1}$ that maps the sphere S^d minus the north pole to the paraboloid $\mathcal{P}$, and this maps satisfies the equation

$$l = \Theta \circ \tau_N.$$

The map Θ is given by

$$\begin{aligned} z_i &= \frac{x_i}{1 - x_{d+1}}, \qquad 1 \leq i \leq d \\ z_{d+1} &= \frac{x_{d+1} + 1}{1 - x_{d+1}}. \end{aligned}$$

Observe that Θ is actually a partial function which is undefined on the hyperplane H_{d+1} tangent to S^d at the north pole, and that its first d component are identical to those of the stereographic projection! Then, we immediately find that

$$\begin{aligned} x_i &= \frac{2z_i}{1 + z_{d+1}}, \qquad 1 \leq i \leq d \\ x_{d+1} &= \frac{z_{d+1} - 1}{1 + z_{d+1}}. \end{aligned}$$

Consequently, Θ is a bijection between $\mathbb{E}^{d+1} - H_{d+1}$ and $\mathbb{E}^{d+1} - H_{d+1}(-1)$, where $H_{d+1}(-1)$ is the hyperplane of equation $x_{d+1} = -1$. As we said earlier, Θ maps the sphere S^d minus the north pole to the paraboloid $\mathcal{P}$, (see Figure 13.13),and

$$l = \Theta \circ \tau_N.$$

What this means is that if we think of the inverse stereographic projection τ_N as a lifting of points in $\mathbb{E}^d$ to the sphere S^d, then lifting points from $\mathbb{E}^d$ to S^d and then mapping $S^d - \{N\}$ to $\mathcal{P}$ by applying Θ is equivalent to lifting points from $\mathbb{E}^d$ to the paraboloid $\mathcal{P}$ using l.

It would be tempting to define the Voronoi diagram $\mathcal{V}or(P)$ as the central projection of the polar dual $\mathcal{D}(P)^*$ of $\mathcal{D}(P)$. However, we have to be careful because Θ does not map all convex polyhedra to convex polyhedra. In particular, Θ is not well-defined on any face of $\mathcal{D}(P)^*$ intersecting the hyperplane H_{d+1} (of equation $x_{d+1} = 1$). Fortunately, we can circumvent these difficulties by using the concept of a projective polyhedron introduced in Chapter 12 and defining a projective version θ of Θ which is a total function. We can also define projective versions of σ_N, τ_N, l, and π_N, to prove that the Voronoi diagram of P is indeed obtained from a suitable projection of the polar dual of $\mathcal{D}(P)$ (actually, a projective version of $\mathcal{D}(P)$).

In summary, Voronoi diagrams, Delaunay Triangulations, and their properties, can also be nicely explained using inverse stereographic projection and the central projection from N, but a rigorous justification of why this "works" is not as simple as it might appear.

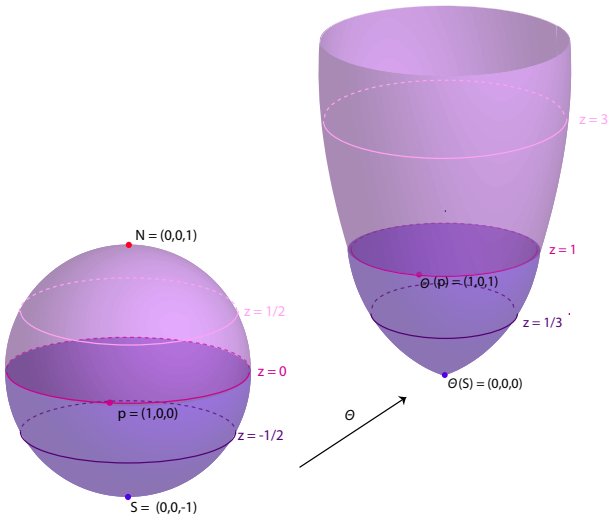


Figure 13.13: The projective transformation Θ which maps $S^2 - \{N\}$ to $z = x^2 + y^2$.

The advantage of stereographic projection over the lifting onto a paraboloid is that the (d-)sphere is compact. Since the stereographic projection and its inverse map $(d-1)$-spheres to $(d-1)$-spheres (or hyperplanes), all the crucial properties of Delaunay triangulations are preserved. The purpose of this section is to establish the properties of stereographic projection (and its inverse) that will be needed in Section 13.7.

Recall that the *d-sphere* $S^d \subseteq \mathbb{E}^{d+1}$ is given by

$$S^d = \{(x_1, \ldots, x_{d+1}) \in \mathbb{E}^{d+1} \mid x_1^2 + \cdots + x_d^2 + x_{d+1}^2 = 1\}.$$

It will be convenient to write a point $(x_1, \ldots, x_{d+1}) \in \mathbb{E}^{d+1}$ as $z = (x, x_{d+1})$, with $x = (x_1, \ldots, x_d)$. We denote $N = (0, \ldots, 0, 1)$ (with d zeros) as $(\mathbf{0}, 1)$ and call it the *north pole*, and $S = (0, \ldots, 0, -1)$ (with d zeros) as $(\mathbf{0}, -1)$ and call it the *south pole*. We also write $\|z\| = (x_1^2 + \cdots + x_{d+1}^2)^{\frac{1}{2}} = (\|x\|^2 + x_{d+1}^2)^{\frac{1}{2}}$ (with $\|x\| = (x_1^2 + \cdots + x_d^2)^{\frac{1}{2}}$). With these notations,

$$S^d = \{(x, x_{d+1}) \in \mathbb{E}^{d+1} \mid \|x\|^2 + x_{d+1}^2 = 1\}.$$

The *stereographic projection from the north pole* $\sigma_N \colon (S^d - \{N\}) \to \mathbb{E}^d$ is the restriction to S^d of the *central projection* $\pi_N \colon (\mathbb{E}^{d+1} - H_{d+1}) \to \mathbb{E}^d$ from N onto the hyperplane $H_{d+1}(0) \cong \mathbb{E}^d$ of equation $x_{d+1} = 0$; that is, $M \mapsto \pi_N(M)$ where $\pi_N(M)$ is the intersection of the line $\langle N, M\rangle$ through N and M with $H_{d+1}(0)$. Since the line through N and $M = (x, x_{d+1})$ is given parametrically by

$$\langle N, M\rangle = \{(1-\lambda)(\mathbf{0}, 1) + \lambda(x, x_{d+1}) \mid \lambda \in \mathbb{R}\},$$

the intersection $\pi_N(M)$ of this line with the hyperplane $x_{d+1} = 0$ corresponds to the value of λ such that

$$(1 - \lambda) + \lambda x_{d+1} = 0,$$

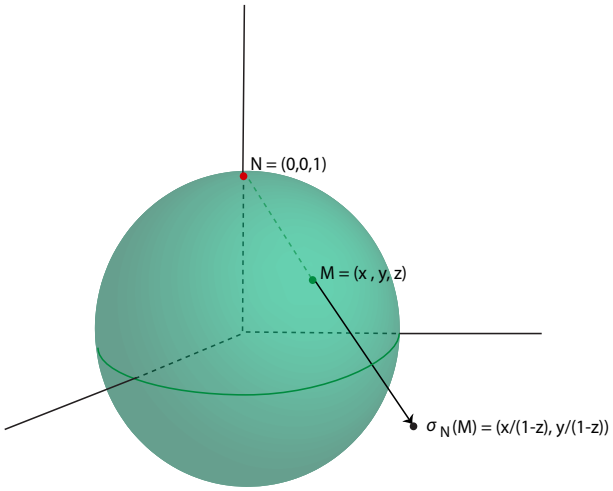


Figure 13.14: The stereographic projection $\sigma_N \colon (S^2 - \{N\}) \to \mathbb{E}^2$.

that is,

$$\lambda = \frac{1}{1 - x_{d+1}}.$$

Therefore, the coordinates of $\pi_N(M)$, with $M = (x, x_{d+1})$, are given by

$$\pi_N(x, x_{d+1}) = \frac{x}{1 - x_{d+1}}.$$

See Figure 13.14. The central projection π_N is undefined on the hyperplane H_{d+1} of equation $x_{d+1} = 1$, and the stereographic projection σ_N from the north pole, which is the restriction of π_N to the sphere S^d, is undefined at the north pole.

Let us find the inverse $\tau_N = \sigma_N^{-1}(P)$ of any $P \in H_{d+1}(0) \cong \mathbb{E}^d$. This time, $\tau_N(P)$ is the intersection of the line $\langle N, P\rangle$ through $P \in H_{d+1}(0)$ and N with the sphere S^d. Since the line through N and $P = (x, 0)$ is given parametrically by

$$\langle N, P\rangle = \{(1 - \lambda)(\mathbf{0}, 1) + \lambda(x, 0) \mid \lambda \in \mathbb{R}\},$$

the intersection $\tau_N(P)$ of this line with the sphere S^d corresponds to the nonzero value of λ such that

$$\lambda^2 \|x\|^2 + (1 - \lambda)^2 = 1,$$

that is

$$\lambda(\lambda(\|x\|^2 + 1) - 2) = 0.$$

Thus, we get

$$\lambda = \frac{2}{\|x\|^2 + 1},$$

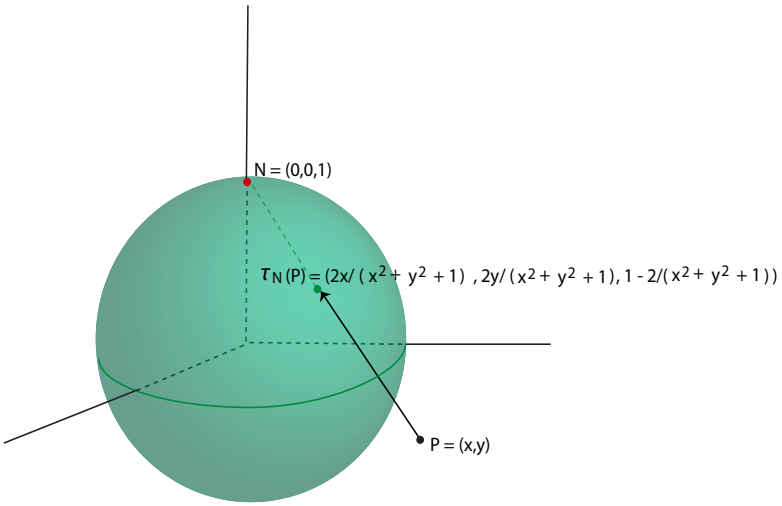


Figure 13.15: The inverse stereographic projection $\tau_N \colon \mathbb{E}^2 \to (S^2 - \{N\})$.

from which we get

$$\begin{aligned} \tau_N(x) &= \left(\frac{2x}{\|x\|^2 + 1}, 1 - \frac{2}{\|x\|^2 + 1} \right) \\ &= \left(\frac{2x}{\|x\|^2 + 1}, \frac{\|x\|^2 - 1}{\|x\|^2 + 1} \right). \end{aligned}$$

See Figure 13.15.

We leave it as an exercise to the reader to verify that $\tau_N \circ \sigma_N = \mathrm{id}$ and $\sigma_N \circ \tau_N = \mathrm{id}$. We can also define the *stereographic projection from the south pole* $\sigma_S \colon (S^d - \{S\}) \to \mathbb{E}^d$, and its inverse τ_S. Again, the computations are left as a simple exercise to the reader. The above computations are summarized in the following definition:

Definition 13.5. The *central projection* $\pi_N \colon (\mathbb{E}^{d+1} - H_{d+1}) \to \mathbb{E}^d$ from N onto the hyperplane $H_{d+1}(0) \cong \mathbb{E}^d$ of equation $x_{d+1} = 0$ is given by

$$\pi_N(x, x_{d+1}) = \frac{x}{1 - x_{d+1}}, \qquad (x_{d+1} \neq 1).$$

The *stereographic projection from the north pole* $\sigma_N \colon (S^d - \{N\}) \to \mathbb{E}^d$ is the restriction of π_N to the sphere S^d. The inverse of σ_N, denoted $\tau_N \colon \mathbb{E}^d \to (S^d - \{N\})$ and called *inverse stereographic projection from the north pole*, is given by

$$\tau_N(x) = \left(\frac{2x}{\|x\|^2 + 1}, \frac{\|x\|^2 - 1}{\|x\|^2 + 1} \right).$$

Remark: An *inversion of center C and power $\rho > 0$* is a geometric transformation $f\colon (\mathbb{E}^{d+1} - \{C\}) \to \mathbb{E}^{d+1}$ defined so that for any $M \neq C$, the points C, M, and $f(M)$ are collinear, and

$$\|\mathbf{CM}\|\|\mathbf{Cf(M)}\| = \rho.$$

Equivalently, $f(M)$ is given by

$$f(M) = C + \frac{\rho}{\|\mathbf{CM}\|^2}\,\mathbf{CM}.$$

Clearly, $f \circ f = \mathrm{id}$ on $\mathbb{E}^{d+1} - \{C\}$, so f is invertible and the reader will check that if we pick the center of inversion to be the north pole and if we set $\rho = 2$, then the coordinates of $f(M)$ are given by

$$\begin{aligned} y_i &= \frac{2x_i}{x_1^2 + \cdots + x_d^2 + x_{d+1}^2 - 2x_{d+1} + 1}, \qquad 1 \leq i \leq d \\ y_{d+1} &= \frac{x_1^2 + \cdots + x_d^2 + x_{d+1}^2 - 1}{x_1^2 + \cdots + x_d^2 + x_{d+1}^2 - 2x_{d+1} + 1}, \end{aligned}$$

where $(x_1, \ldots, x_{d+1})$ are the coordinates of M. In particular, if we restrict our inversion to the unit sphere S^d, as $x_1^2 + \cdots + x_d^2 + x_{d+1}^2 = 1$, we get

$$\begin{aligned} y_i &= \frac{x_i}{1 - x_{d+1}}, \qquad 1 \leq i \leq d \\ y_{d+1} &= 0, \end{aligned}$$

which means that our inversion restricted to S^d is simply the stereographic projection σ_N (and the inverse of our inversion restricted to the hyperplane $x_{d+1} = 0$ is the inverse stereographic projection τ_N).

We will now show that the image of any $(d-1)$-sphere S on S^d not passing through the north pole, that is, the intersection $S = S^d \cap H$ of S^d with any hyperplane H not passing through N, is a $(d-1)$-sphere. Here, we are assuming that S has positive radius, that is, H is *not* tangent to S^d.

Assume that H is given by

$$a_1x_1 + \cdots + a_dx_d + a_{d+1}x_{d+1} + b = 0.$$

Since $N \notin H$, we must have $a_{d+1} + b \neq 0$. For any $(x, x_{d+1}) \in S^d$, write $\sigma_N(x, x_{d+1}) = X$. Since

$$X = \frac{x}{1 - x_{d+1}},$$

we get $x = X(1 - x_{d+1})$, and using the fact that (x, x_{d+1}) also belongs to H we will express x_{d+1} in terms of X and then find an equation for X which will show that X belongs to a $(d-1)$-sphere. Indeed, $(x, x_{d+1}) \in H$ implies that

$$\sum_{i=1}^{d} a_iX_i(1 - x_{d+1}) + a_{d+1}x_{d+1} + b = 0,$$

that is,

$$\sum_{i=1}^{d} a_i X_i + (a_{d+1} - \sum_{j=1}^{d} a_j X_j) x_{d+1} + b = 0.$$

If $\sum_{j=1}^{d} a_j X_j = a_{d+1}$, then $a_{d+1} + b = 0$, which is impossible. Therefore, we get

$$x_{d+1} = \frac{-b - \sum_{i=1}^{d} a_i X_i}{a_{d+1} - \sum_{i=1}^{d} a_i X_i},$$

and so

$$1 - x_{d+1} = \frac{a_{d+1} + b}{a_{d+1} - \sum_{i=1}^{d} a_i X_i}.$$

Plugging $x = X(1 - x_{d+1})$ in the equation $\|x\|^2 + x_{d+1}^d = 1$ of S^d, we get

$$(1 - x_{d+1})^2 \|X\|^2 + x_{d+1}^2 = 1,$$

and replacing x_{d+1} and $1 - x_{d+1}$ by their expression in terms of X, we get

$$(a_{d+1} + b)^2 \|X\|^2 + (-b - \sum_{i=1}^{d} a_i X_i)^2 = (a_{d+1} - \sum_{i=1}^{d} a_i X_i)^2,$$

that is,

$$\begin{aligned} (a_{d+1} + b)^2 \|X\|^2 &= (a_{d+1} - \sum_{i=1}^{d} a_i X_i)^2 - (b + \sum_{i=1}^{d} a_i X_i)^2 \\ &= (a_{d+1} + b)(a_{d+1} - b - 2\sum_{i=1}^{d} a_i X_i), \end{aligned}$$

which yields

$$(a_{d+1} + b)^2 \|X\|^2 + 2(a_{d+1} + b)(\sum_{i=1}^{d} a_i X_i) = (a_{d+1} + b)(a_{d+1} - b),$$

that is,

$$\|X\|^2 + 2\sum_{i=1}^{d} \frac{a_i}{a_{d+1} + b} X_i - \frac{a_{d+1} - b}{a_{d+1} + b} = 0,$$

which is indeed the equation of a $(d-1)$-sphere in $\mathbb{E}^d$. By "completing the square," the above equation can be written as

$$\sum_{i=1}^{d} \left(X_i + \frac{a_i}{a_{d+1} + b} \right)^2 - \sum_{i=1}^{d} \frac{a_i^2}{(a_{d+1} + b)^2} - \frac{a_{d+1} - b}{a_{d+1} + b} = 0,$$

which yields

$$\sum_{i=1}^{d}\left(X_i + \frac{a_i}{a_{d+1}+b}\right)^2 = \frac{\sum_{i=1}^{d} a_i^2 + (a_{d+1}-b)(a_{d+1}+b)}{(a_{d+1}+b)^2},$$

that is,

$$\sum_{i=1}^{d}\left(X_i + \frac{a_i}{a_{d+1}+b}\right)^2 = \frac{\sum_{i=1}^{d+1} a_i^2 - b^2}{(a_{d+1}+b)^2}. \qquad (*)$$

However, the distance from the origin to the hyperplane H of equation

$$a_1x_1 + \cdots + a_dx_d + a_{d+1}x_{d+1} + b = 0$$

is

$$\delta = \frac{|b|}{\left(\sum_{i=1}^{d+1} a_i^2\right)^{1/2}},$$

and since we are assuming that H intersects the unit sphere S^d in a sphere of positive radius we must have $\delta < 1$, so

$$b^2 < \sum_{i=1}^{d+1} a_i^2,$$

and $(*)$ is indeed the equation of a real sphere (its radius is positive). Therefore, when $N \notin H$, the image of $S = S^d \cap H$ by σ_N is a $(d-1)$-sphere in $H_{d+1}(0) = \mathbb{E}^d$. See Figure 13.16.

If the hyperplane H contains the north pole, then $a_{d+1} + b = 0$, in which case, for every $(x, x_{d+1}) \in S^d \cap H$, we have

$$\sum_{i=1}^{d} a_ix_i + a_{d+1}x_{d+1} - a_{d+1} = 0,$$

that is,

$$\sum_{i=1}^{d} a_ix_i - a_{d+1}(1 - x_{d+1}) = 0,$$

and except for the north pole, we have

$$\sum_{i=1}^{d} a_i\frac{x_i}{1 - x_{d+1}} - a_{d+1} = 0,$$

which shows that

$$\sum_{i=1}^{d} a_iX_i - a_{d+1} = 0,$$

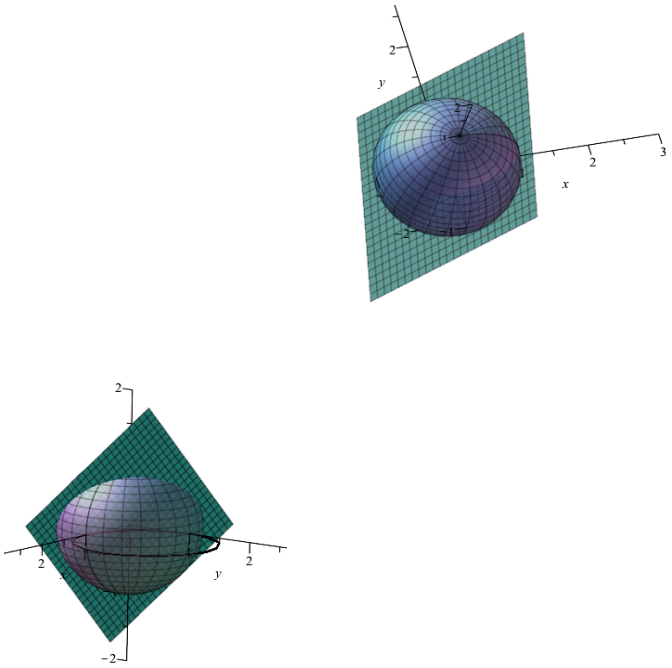


Figure 13.16: Two views of the plane $-x-y+z=0$ intersecting S^2. The bottom figure shows the stereographic projection of the intersection, namely the circle $(x-1)^2+(y-1)^2=1$.

the intersection of the hyperplanes H and $H_{d+1}(0)$. Therefore, the image of $S^d \cap H$ by σ_N is the hyperplane in $\mathbb{E}^d$ which is the intersection of H with $H_{d+1}(0)$. See Figure 13.17.

We will also prove that τ_N maps $(d-1)$-spheres in $H_{d+1}(0)$ to $(d-1)$-spheres on S^d not passing through the north pole. Assume that $X \in \mathbb{E}^d$ belongs to the $(d-1)$-sphere of equation

$$\sum_{i=1}^{d} X_i^2 + \sum_{j=1}^{d} a_j X_j + b = 0.$$

For any $(X,0) \in H_{d+1}(0)$, we know that $(x, x_{d+1}) = \tau_N(X)$ is given by

$$(x, x_{d+1}) = \left(\frac{2X}{\|X\|^2 + 1}, \frac{\|X\|^2 - 1}{\|X\|^2 + 1} \right).$$

Using the equation of the $(d-1)$-sphere, we get

$$x = \frac{2X}{-b + 1 - \sum_{j=1}^{d} a_j X_j}$$

and

$$x_{d+1} = \frac{-b - 1 - \sum_{j=1}^{d} a_j X_j}{-b + 1 - \sum_{j=1}^{d} a_j X_j}.$$

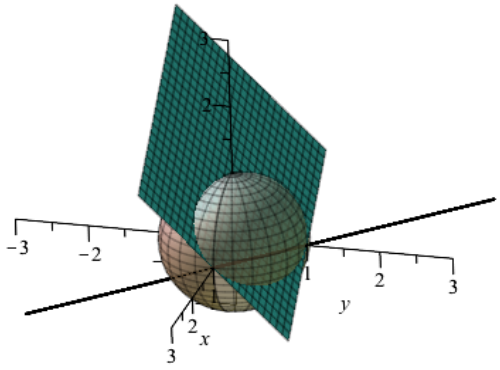


Figure 13.17: The plane $x + y + z - 1 = 0$ intersecting S^2, along with the stereographic projection of the intersection, namely the line $x + y = 1$.

Then, we get

$$\sum_{i=1}^{d} a_i x_i = \frac{2\sum_{j=1}^{d} a_j X_j}{-b+1-\sum_{j=1}^{d} a_j X_j},$$

which yields

$$(-b+1)(\sum_{i=1}^{d} a_i x_i) - (\sum_{i=1}^{d} a_i x_i)(\sum_{j=1}^{d} a_j X_j) = 2\sum_{j=1}^{d} a_j X_j.$$

From the above, we get

$$\sum_{i=1}^{d} a_i X_i = \frac{(-b+1)(\sum_{i=1}^{d} a_i x_i)}{\sum_{i=1}^{d} a_i x_i + 2}.$$

Plugging this expression in the formula for x_{d+1} above, we get

$$x_{d+1} = \frac{-b-1-\sum_{i=1}^{d} a_i x_i}{-b+1},$$

which yields

$$\sum_{i=1}^{d} a_i x_i + (-b+1)x_{d+1} + (b+1) = 0,$$

the equation of a hyperplane H not passing through the north pole. Therefore, the image of a $(d-1)$-sphere in $H_{d+1}(0)$ is indeed the intersection $H \cap S^d$ of S^d with a hyperplane not passing through N, that is, a $(d-1)$-sphere on S^d.

Given any hyperplane H' in $H_{d+1}(0) = \mathbb{E}^d$, say of equation

$$\sum_{i=1}^{d} a_i X_i + b = 0,$$

the image of H' under τ_N is a $(d-1)$-sphere on S^d, the intersection of S^d with the hyperplane H passing through N and determined as follows: For any $(X,0) \in H_{d+1}(0)$, if $\tau_N(X) = (x, x_{d+1})$, then

$$X = \frac{x}{1 - x_{d+1}},$$

and so (x, x_{d+1}) satisfies the equation

$$\sum_{i=1}^{d} a_i x_i + b(1 - x_{d+1}) = 0,$$

that is,

$$\sum_{i=1}^{d} a_i x_i - b x_{d+1} + b = 0,$$

which is indeed the equation of a hyperplane H passing through N. We summarize all this in the following proposition:

Proposition 13.4. *The stereographic projection $\sigma_N \colon (S^d - \{N\}) \to \mathbb{E}^d$ induces a bijection between the set of $(d-1)$-spheres on S^d and the union of the set of $(d-1)$-spheres in $\mathbb{E}^d$ with the set of hyperplanes in $\mathbb{E}^d$; every $(d-1)$-sphere on S^d not passing through the north pole is mapped to a $(d-1)$-sphere in $\mathbb{E}^d$, and every $(d-1)$-sphere on S^d passing through the north pole is mapped to a hyperplane in $\mathbb{E}^d$. In fact, σ_N maps the $(d-1)$-sphere on S^d determined by the hyperplane*

$$a_1 x_1 + \cdots + a_d x_d + a_{d+1} x_{d+1} + b = 0$$

not passing through the north pole ($a_{d+1} + b \neq 0$) to the $(d-1)$-sphere

$$\sum_{i=1}^{d} \left(X_i + \frac{a_i}{a_{d+1} + b} \right)^2 = \frac{\sum_{i=1}^{d+1} a_i^2 - b^2}{(a_{d+1} + b)^2},$$

and the $(d-1)$-sphere on S^d determined by the hyperplane

$$\sum_{i=1}^{d} a_i x_i + a_{d+1} x_{d+1} - a_{d+1} = 0$$

through the north pole to the hyperplane

$$\sum_{i=1}^{d} a_i X_i - a_{d+1} = 0;$$

the map $\tau_N = \sigma_N^{-1}$ maps the $(d-1)$-sphere

$$\sum_{i=1}^{d} X_i^2 + \sum_{j=1}^{d} a_j X_j + b = 0$$

to the $(d-1)$-sphere on S^d determined by the hyperplane

$$\sum_{i=1}^{d} a_i x_i + (-b+1)x_{d+1} + (b+1) = 0$$

not passing through the north pole, and the hyperplane

$$\sum_{i=1}^{d} a_i X_i + b = 0$$

to the $(d-1)$-sphere on S^d determined by the hyperplane

$$\sum_{i=1}^{d} a_i x_i - b x_{d+1} + b = 0$$

through the north pole.

Proposition 13.4 raises a natural question: What do the hyperplanes H in $\mathbb{E}^{d+1}$ that do not intersect S^d correspond to, if they correspond to anything at all?

The first thing to observe is that the geometric definition of the stereographic projection and its inverse makes it clear that the hyperplanes corresponding to $(d-1)$-spheres in $\mathbb{E}^d$ (by τ_N) do intersect S^d. Now, when we write the equation of a $(d-1)$-sphere S, say

$$\sum_{i=1}^{d} X_i^2 + \sum_{i=1}^{d} a_i X_i + b = 0,$$

we are implicitly assuming a condition on the a_i's and b that ensures that S is not the empty sphere, that is, that its radius R is positive (or zero). By "completing the square," the above equation can be rewritten as

$$\sum_{i=1}^{d} \left(X_i + \frac{a_i}{2} \right)^2 = \frac{1}{4} \sum_{i=1}^{d} a_i^2 - b,$$

and so the radius R of our sphere is given by

$$R^2 = \frac{1}{4} \sum_{i=1}^{d} a_i^2 - b$$

whereas its center is the point $c = -\frac{1}{2}(a_1, \ldots, a_d)$. Thus, our sphere is a "real" sphere of positive radius iff

$$\sum_{i=1}^{d} a_i^2 > 4b,$$

or a single point, $c = -\frac{1}{2}(a_1, \ldots, a_d)$, iff $\sum_{i=1}^{d} a_i^2 = 4b$.

What happens when

$$\sum_{i=1}^{d} a_i^2 < 4b?$$

In this case, if we allow "complex points," that is, if we consider solutions of our equation

$$\sum_{i=1}^{d} X_i^2 + \sum_{i=1}^{d} a_i X_i + b = 0$$

over $\mathbb{C}^d$, then we get a "complex" sphere of (pure) imaginary radius $\frac{i}{2}\sqrt{4b - \sum_{i=1}^{d} a_i^2}$. The funny thing is that our computations carry over unchanged and the image of the complex sphere S is still the intersection of the complex sphere S^d with the hyperplane H given

$$\sum_{i=1}^{d} a_i x_i + (-b+1)x_{d+1} + (b+1) = 0.$$

However, this time, even though H does not have any "real" intersection points with S^d, we can show that it does intersect the "complex sphere,"

$$S^d = \{(z_1, \ldots, z_{d+1}) \in \mathbb{C}^{d+1} \mid z_1^2 + \cdots + z_{d+1}^2 = 1\}$$

in a nonempty set of points in $\mathbb{C}^{d+1}$.

It follows from all this that σ_N and τ_N establish a bijection between the set of all hyperplanes in $\mathbb{E}^{d+1}$ minus the hyperplane H_{d+1} (of equation $x_{d+1} = 1$) tangent to S^d at the north pole, with the union of four sets:

(1) The set of all (real) $(d-1)$-spheres of positive radius; see Figure 13.16.

(2) The set of all (complex) $(d-1)$-spheres of imaginary radius;

(3) The set of all hyperplanes in $\mathbb{E}^d$; see Figure 13.17.

(4) The set of all points of $\mathbb{E}^d$ (viewed as spheres of radius 0); see Figure 13.18.

Moreover, Set (1) corresponds to the hyperplanes that intersect the interior of S^d and do not pass through the north pole; Set (2) corresponds to the hyperplanes that do not intersect S^d; Set (3) corresponds to the hyperplanes that pass through the north pole minus the tangent hyperplane at the north pole; and Set (4) corresponds to the hyperplanes that are tangent to S^d, minus the tangent hyperplane at the north pole.

It is convenient to add the "point at infinity" ∞ to $\mathbb{E}^d$, because then the above bijection can be extended to map the tangent hyperplane at the north pole to ∞. The union of these four sets (with ∞ added) is called the *set of generalized spheres*, sometimes denoted $\mathcal{S}(\mathbb{E}^d)$.

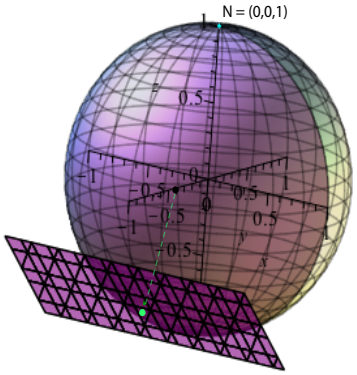


Figure 13.18: The plane $-2y - 2z - 2\sqrt{2} = 0$ tangent to the point $(0, -1/\sqrt{(}2), -1/\sqrt{2})$, along with its corresponding stereographic projection, the point $(0, -1/(\sqrt{2}+1), 0)$.

This is a fairly complicated space. For one thing, topologically $\mathcal{S}(\mathbb{E}^d)$ is homeomorphic to the projective space $\mathbb{P}^{d+1}$ with one point removed (the point corresponding to the "hyperplane at infinity"), and this is not a simple space. We can get a slightly more concrete "'picture" of $\mathcal{S}(\mathbb{E}^d)$ by looking at the polars of the hyperplanes w.r.t. S^d. Then, the "real" spheres correspond to the points strictly outside S^d which do not belong to the tangent hyperplane at the north pole, (i.e. Figure 13.19); the complex spheres correspond to the points in the interior of S^d; the points of $\mathbb{E}^d \cup \{\infty\}$ correspond to the points on S^d, (i.e. Figure 13.18); the hyperplanes in $\mathbb{E}^d$ correspond to the points in the tangent hyperplane at the north pole expect for the north pole, (i.e. Figure 13.20). Unfortunately, the poles of hyperplanes through the origin are undefined. This can be fixed by embedding $\mathbb{E}^{d+1}$ in its projective completion $\mathbb{P}^{d+1}$, but we will not go into this.

There are other ways of dealing rigorously with the set of generalized spheres. One method described by Boissonnat [12] is to use the embedding where the sphere S of equation

$$\sum_{i=1}^{d} X_i^2 - 2\sum_{i=1}^{d} a_i X_i + b = 0$$

is mapped to the point

$$\varphi(S) = (a_1, \ldots, a_d, b) \in \mathbb{E}^{d+1}.$$

This gives us another way of dealing with the sets of type (1), (2), and (4) described earlier. Now, by a previous computation we know that

$$b = \sum_{i=1}^{d} a_i^2 - R^2,$$

where $c = (a_1, \ldots, a_d)$ is the center of S and R is its radius.

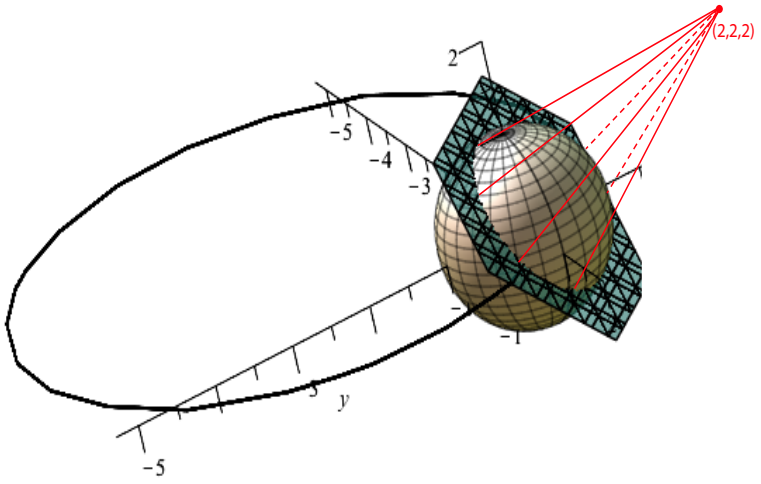


Figure 13.19: The plane $2x + 2y + 2z = 1$ with its dual $(2, 2, 2)$. Also shown is the stereographic projection of the intersection, namely the circle $(x + 2)^2 + (y + 2)^2 = 11$.

The quantity $\sum_{i=1}^{d} a_i^2 - R^2$ is known as the *power* of the origin w.r.t. S. In general, the *power* of a point $X \in \mathbb{E}^d$ is defined as $\rho(X) = \|\mathbf{cX}\|^2 - R^2$, which, after a moment of thought, is just

$$\rho(X) = \sum_{i=1}^{d} X_i^2 - 2\sum_{i=1}^{d} a_i X_i + b.$$

Now, since points correspond to spheres of radius 0, we see that the image of the point $X = (X_1, \ldots, X_d)$ is

$$l(X) = (X_1, \ldots, X_d, \sum_{i=1}^{d} X_i^2).$$

Thus, in this model, points of $\mathbb{E}^d$ are lifted to the paraboloid $\mathcal{P} \subseteq \mathbb{E}^{d+1}$ of equation

$$x_{d+1} = \sum_{i=1}^{d} x_i^2.$$

Actually, this method does not deal with hyperplanes but it is possible to do so. The trick is to consider equations of a slightly more general form that capture both spheres and hyperplanes, namely, equations of the form

$$c\sum_{i=1}^{d} X_i^2 + \sum_{i=1}^{d} a_i X_i + b = 0.$$

Indeed, when $c = 0$, we do get a hyperplane! Now, to carry out this method we really need to consider equations up to a nonzero scalars, that is, we consider the projective space

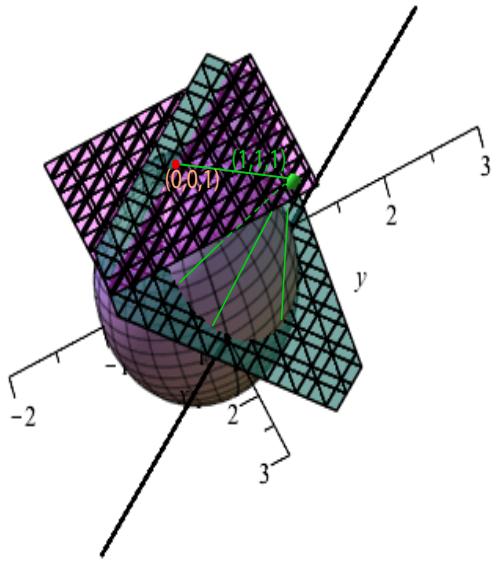


Figure 13.20: The plane $x+y+z=1$ with its dual $(1,1,1)$. Also shown is the stereographic projection of the intersection, namely the line $x+y=1$.

$\mathbb{P}(\widehat{S}(\mathbb{E}^d))$ associated with the vector space $\widehat{S}(\mathbb{E}^d)$ consisting of the above equations. Then, it turns out that the quantity

$$\varrho(a,b,c) = \frac{1}{4}(\sum_{i=1}^{d} a_i^2 - 4bc)$$

(with $a=(a_1,\ldots,a_d)$) defines a quadratic form on $\widehat{S}(\mathbb{E}^d)$ whose corresponding bilinear form

$$\rho((a,b,c),(a',b',c')) = \frac{1}{4}(\sum_{i=1}^{d} a_i a_i' - 2bc' - 2b'c)$$

has a natural interpretation (with $a=(a_1,\ldots,a_d)$ and $a'=(a_1',\ldots,a_d')$). Indeed, orthogonality with respect to ρ (that is, when $\rho((a,b,c),(a',b',c'))=0$) says that the corresponding spheres defined by (a,b,c) and (a',b',c') are orthogonal, that the corresponding hyperplanes defined by $(a,b,0)$ and $(a',b',0)$ are orthogonal, *etc.* The reader who wants to read more about this approach should consult Berger (Volume II) [8].

13.6 Relating Lifting to a Paraboloid and Lifting to a Sphere

We explained in Section 13.5 that there is a simple relationship between the lifting onto a paraboloid and the lifting onto S^d using the inverse stereographic projection map because

the sphere and the paraboloid are projectively equivalent, as we showed for S^2 in Section 12.1.

We defined the map Θ given by

$$\begin{aligned} z_i &= \frac{x_i}{1 - x_{d+1}}, \qquad 1 \leq i \leq d \\ z_{d+1} &= \frac{x_{d+1} + 1}{1 - x_{d+1}}, \end{aligned}$$

and showed that Θ is a bijection between $\mathbb{E}^{d+1} - H_{d+1}$ and $\mathbb{E}^{d+1} - H_{d+1}(-1)$, where $H_{d+1}(-1)$ is the hyperplane of equation $x_{d+1} = -1$. We will show a little later that Θ maps the sphere S^d minus the north pole to the paraboloid $\mathcal{P}$, and satisfies the equation

$$l = \Theta \circ \tau_N.$$

The fact that Θ is undefined on the hyperplane H_{d+1} is not a problem as far as mapping the sphere to the paraboloid because the north pole is the only point that does not have an image. However, later on when we consider the Voronoi polyhedron $\mathcal{V}(P)$ of a lifted set of points P, we will have more serious problems because in general, such a polyhedron intersects both hyperplanes H_{d+1} and $H_{d+1}(-1)$. This means that Θ will not be well-defined on the whole of $\mathcal{V}(P)$ nor will it be surjective on its image. To remedy this difficulty, we work with projective completions. Basically, this amounts to chasing denominators and homogenizing equations, but we also have to be careful in dealing with convexity, and this is where the projective polyhedra (studied in Section 12.2) will come handy.

So, let us consider the projective completion of the sphere $\widetilde{S^d} \subseteq \mathbb{P}^{d+1}$ given by the equation

$$\sum_{i=1}^{d+1} x_i^2 = x_{d+2}^2,$$

and the projective completion of the paraboloid $\widetilde{\mathcal{P}} \subseteq \mathbb{P}^{d+1}$ given by the equation

$$x_{d+1}x_{d+2} = \sum_{i=1}^{d} x_i^2.$$

Definition 13.6. Let $\theta\colon \mathbb{P}^{d+1} \to \mathbb{P}^{d+1}$ be the projectivity induced by the linear map $\widehat{\theta}\colon \mathbb{R}^{d+2} \to \mathbb{R}^{d+2}$ given by

$$\begin{aligned} z_i &= x_i, \qquad 1 \leq i \leq d \\ z_{d+1} &= x_{d+1} + x_{d+2} \\ z_{d+2} &= x_{d+2} - x_{d+1}, \end{aligned}$$

whose inverse is given by

$$\begin{aligned} x_i &= z_i, && 1 \le i \le d \\ x_{d+1} &= \frac{z_{d+1} - z_{d+2}}{2} \\ x_{d+2} &= \frac{z_{d+1} + z_{d+2}}{2}. \end{aligned}$$

The map θ is a projective version of Θ, but is better behaved because it is a total function. If we plug these formulae in the equation of $\widetilde{S^d}$, we get

$$4(\sum_{i=1}^{d} z_i^2) + (z_{d+1} - z_{d+2})^2 = (z_{d+1} + z_{d+2})^2,$$

which simplifies to

$$z_{d+1} z_{d+2} = \sum_{i=1}^{d} z_i^2.$$

Therefore, $\theta(\widetilde{S^d}) = \widetilde{\mathcal{P}}$, that is, θ maps the projective completion of the sphere to the projective completion of the paraboloid. Observe that the projective north pole $\widetilde{N} = (0\colon \cdots \colon 0\colon 1\colon 1)$ is mapped to the point at infinity $(0\colon \cdots \colon 0\colon 1\colon 0)$.

Recall from Definition 12.5 that for any i, with $1 \le i \le d+1$, the set

$$U_i = \{(x_1\colon \cdots \colon x_{d+1}) \in \mathbb{P}^d \mid x_i \neq 0\}$$

is a subset of $\mathbb{P}^d$ called an *affine patch* of $\mathbb{P}^d$. We have a bijection $\varphi_i\colon U_i \to \mathbb{R}^d$ between U_i and $\mathbb{R}^d$ given by

$$\varphi_i\colon (x_1\colon \cdots \colon x_{d+1}) \mapsto \left(\frac{x_1}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_{d+1}}{x_i}\right),$$

with inverse $\psi_i\colon \mathbb{R}^d \to U_i \subseteq \mathbb{P}^d$ given by

$$\psi_i\colon (x_1, \cdots, x_d) \mapsto (x_1\colon \cdots x_{i-1}\colon 1\colon x_i\colon \cdots \colon x_d).$$

The map Θ is the restriction of θ to the affine patch U_{d+1}, and as such, it can be fruitfully described as the composition of $\widehat{\theta}$ with a suitable projection onto $\mathbb{E}^{d+1}$. For this, as we have done before, we identify $\mathbb{E}^{d+1}$ with the hyperplane $H_{d+2} \subseteq \mathbb{E}^{d+2}$ of equation $x_{d+2} = 1$ (using the injection, $i_{d+2}\colon \mathbb{E}^{d+1} \to \mathbb{E}^{d+2}$, where $i_j\colon \mathbb{E}^{d+1} \to \mathbb{E}^{d+2}$ is the injection given by

$$(x_1, \ldots, x_{d+1}) \mapsto (x_1, \ldots, x_{j-1}, 1, x_{j+1}, \ldots, x_{d+1})$$

for any $(x_1, \ldots, x_{d+1}) \in \mathbb{E}^{d+1}$). For each i, with $1 \le i \le d+2$, let $\pi_i\colon (\mathbb{E}^{d+2} - H_i(0)) \to \mathbb{E}^{d+1}$ be the projection of center $0 \in \mathbb{E}^{d+2}$ onto the hyperplane $H_i \subseteq \mathbb{E}^{d+2}$ of equation $x_i = 1$ ($H_i \cong \mathbb{E}^{d+1}$ and $H_i(0) \subseteq \mathbb{E}^{d+2}$ is the hyperplane of equation $x_i = 0$) given by

$$\pi_i(x_1, \ldots, x_{d+2}) = \left(\frac{x_1}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_{d+2}}{x_i}\right) \qquad (x_i \neq 0).$$

Geometrically, for any $x \notin H_i(0)$, the image $\pi_i(x)$ of x is the intersection of the line through the origin and x with the hyperplane $H_i \subseteq \mathbb{E}^{d+2}$ of equation $x_i = 1$. This is illustrated in Figure 13.21. Observe that the map $\pi_i \colon (\mathbb{E}^{d+2} - H_{d+2}(0)) \to \mathbb{E}^{d+1}$ is an "affine" version of the bijection $\varphi_i \colon U_i \to \mathbb{R}^{d+1}$ of Section 12.1. Then, we have

$$\Theta = \pi_{d+2} \circ \widehat{\theta} \circ i_{d+2}.$$

If we identify H_{d+2} and $\mathbb{E}^{d+1}$, we may write with a slight abuse of notation $\Theta = \pi_{d+2} \circ \widehat{\theta}$.

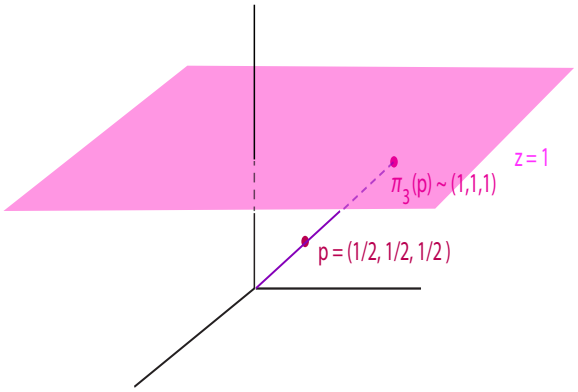


Figure 13.21: The geometric realization of image of $\pi_3(p)$, where $\pi_3 \colon (\mathbb{E}^3 - H_3(0)) \to \mathbb{E}^2$.

We will need some properties of the projection π_{d+2} and of Θ, and for this, let

$$\mathbb{H}^d_+ = \{(x_1, \ldots, x_d) \in \mathbb{E}^d \mid x_d > 0\} \quad \text{and} \quad \mathbb{H}^d_- = \{(x_1, \ldots, x_d) \in \mathbb{E}^d \mid x_d < 0\}.$$

Proposition 13.5. *The maps π_{d+2}, π_N, and Θ have the following properties:*

(1) For every hyperplane H through the origin, $\pi_{d+2}(H)$ is a hyperplane in H_{d+2}. See Figure 13.22.

(2) Given any set of points $\{a_1, \ldots, a_n\} \subseteq \mathbb{E}^{d+2}$, if $\{a_1, \ldots, a_n\}$ is contained in the open half-space above the hyperplane $x_{d+2} = 0$ or $\{a_1, \ldots, a_n\}$ is contained in the open half-space below the hyperplane $x_{d+2} = 0$, then the image by π_{d+2} of the convex hull of the a_i's is the convex hull of the images of these points, that is,

$$\pi_{d+2}(\mathrm{conv}(\{a_1, \ldots, a_n\})) = \mathrm{conv}(\{\pi_{d+2}(a_1), \ldots, \pi_{d+2}(a_n)\}).$$

See Figure 13.23.

(3) Given any set of points $\{a_1, \ldots, a_n\} \subseteq \mathbb{E}^{d+2}$*, if* $\{a_1, \ldots, a_n\}$ *is contained in the open half-space above the hyperplane* $x_{d+2} = 1$ *or* $\{a_1, \ldots, a_n\}$ *is contained in the open half-space below the hyperplane* $x_{d+2} = 1$*, then the image by* π_N *of the convex hull of the* a_i*'s is the convex hull of the images of these points, that is,*

$$\pi_N(\mathrm{conv}(\{a_1, \ldots, a_n\})) = \mathrm{conv}(\{\pi_N(a_1), \ldots, \pi_N(a_n)\}).$$

(4) Given any set of points $\{a_1, \ldots, a_n\} \subseteq \mathbb{E}^{d+1}$*, if* $\{a_1, \ldots, a_n\}$ *is contained in the open half-space above the hyperplane* H_{d+1} *or* $\{a_1, \ldots, a_n\}$ *is contained in the open half-space below* H_{d+1}*, then*

$$\Theta(\mathrm{conv}(\{a_1, \ldots, a_n\})) = \mathrm{conv}(\{\Theta(a_1), \ldots, \Theta(a_n)\}).$$

(5) For any set $S \subseteq \mathbb{E}^{d+1}$*, if* $\mathrm{conv}(S)$ *does not intersect* H_{d+1}*, then*

$$\Theta(\mathrm{conv}(S)) = \mathrm{conv}(\Theta(S)).$$

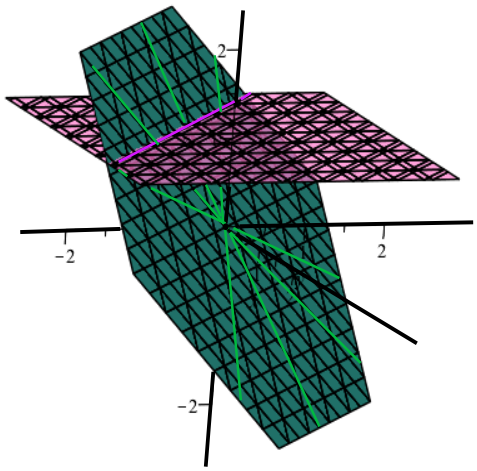


Figure 13.22: The intersection of the teal plane $x + y + z = 0$ with the magenta plane $z = 1$ results in the pink line $x + y = -1$. This line is the also the projection of the teal plane via π_3 as shown by the lime green rays through the origin.

Proof. (1) The image, $\pi_{d+2}(H)$, of a hyperplane H through the origin is the intersection of H with H_{d+2}, which is a hyperplane in H_{d+2}.

(2) This seems fairly clear geometrically but the result fails for arbitrary sets of points, so to be on the safe side, we give an algebraic proof. We will prove the following two facts by induction on $n \geq 1$:

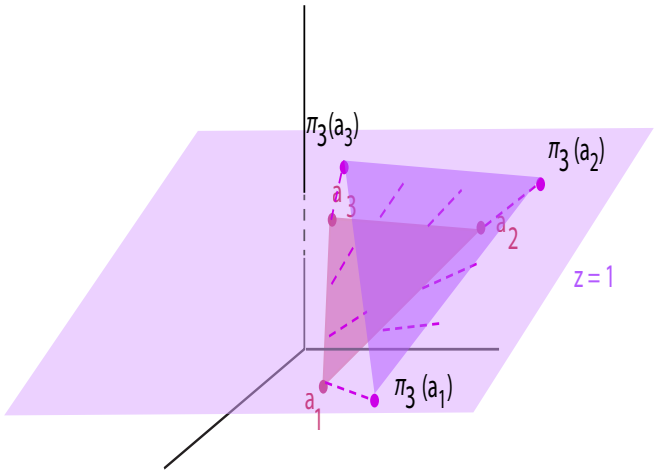


Figure 13.23: The convex hull of $\{a_1, a_2, a_3\}$, namely the dusty rose triangle above $z = 0$ and below $z = 1$, under π_3, is projected to the lavender triangle in the plane $z = 1$.

(1) For all $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ with $\lambda_1 + \cdots + \lambda_n = 1$ and $\lambda_i \geq 0$, for all $a_1, \ldots, a_n \in \mathbb{H}_+^{d+2}$ (resp. $\in \mathbb{H}_-^{d+2}$), there exist some $\mu_1, \ldots, \mu_n \in \mathbb{R}$ with $\mu_1 + \cdots + \mu_n = 1$ and $\mu_i \geq 0$, so that

$$\pi_{d+2}(\lambda_1 a_1 + \cdots + \lambda_n a_n) = \mu_1 \pi_{d+2}(a_1) + \cdots + \mu_n \pi_{d+2}(a_n).$$

(2) For all $\mu_1, \ldots, \mu_n \in \mathbb{R}$ with $\mu_1 + \cdots + \mu_n = 1$ and $\mu_i \geq 0$, for all $a_1, \ldots, a_n \in \mathbb{H}_+^{d+2}$ (resp. $\in \mathbb{H}_-^{d+2}$), there exist some $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ with $\lambda_1 + \cdots + \lambda_n = 1$ and $\lambda_i \geq 0$, so that

$$\pi_{d+2}(\lambda_1 a_1 + \cdots + \lambda_n a_n) = \mu_1 \pi_{d+2}(a_1) + \cdots + \mu_n \pi_{d+2}(a_n).$$

(1) The base case is clear. Let us assume for the moment that we proved (1) for $n = 2$ and consider the induction step for $n \geq 2$. Since $\lambda_1 + \cdots + \lambda_{n+1} = 1$ and $n \geq 2$, there is some i such that $\lambda_i \neq 1$, and without loss of generality, say $\lambda_1 \neq 1$. Then, we can write

$$\lambda_1 a_1 + \cdots + \lambda_{n+1} a_{n+1} = \lambda_1 a_1 + (1 - \lambda_1)\left(\frac{\lambda_2}{1 - \lambda_1} a_2 + \cdots + \frac{\lambda_{n+1}}{1 - \lambda_1} a_{n+1}\right)$$

and since $\lambda_1 + \lambda_2 + \cdots + \lambda_{n+1} = 1$, we have

$$\frac{\lambda_2}{1 - \lambda_1} + \cdots + \frac{\lambda_{n+1}}{1 - \lambda_1} = 1.$$

By the induction hypothesis, for $n = 2$, there exist α_1 with $0 \leq \alpha_1 \leq 1$, such that

$$\begin{aligned}\pi_{d+2}(\lambda_1 a_1 + \cdots + \lambda_{n+1} a_{n+1}) &= \pi_{d+2}\left(\lambda_1 a_1 + (1-\lambda_1)\left(\frac{\lambda_2}{1-\lambda_1} a_2 + \cdots + \frac{\lambda_{n+1}}{1-\lambda_1} a_{n+1}\right)\right) \\ &= (1-\alpha_1)\pi_{d+2}(a_1) + \alpha_1 \pi_{d+2}\left(\frac{\lambda_2}{1-\lambda_1} a_2 + \cdots + \frac{\lambda_{n+1}}{1-\lambda_1} a_{n+1}\right).\end{aligned}$$

Again, by the induction hypothesis (for n), there exist $\beta_2, \ldots, \beta_{n+1}$ with $\beta_2 + \cdots + \beta_{n+1} = 1$ and $\beta_i \geq 0$, so that

$$\pi_{d+2}\left(\frac{\lambda_2}{1-\lambda_1} a_2 + \cdots + \frac{\lambda_{n+1}}{1-\lambda_1} a_{n+1}\right) = \beta_2 \pi_{d+2}(a_2) + \cdots + \beta_{n+1}\pi_{d+2}(a_{n+1}),$$

so we get

$$\begin{aligned}\pi_{d+2}(\lambda_1 a_1 + \cdots + \lambda_{n+1} a_{n+1}) &= (1-\alpha_1)\pi_{d+2}(a_1) + \alpha_1(\beta_2 \pi_{d+2}(a_2) + \cdots + \beta_{n+1}\pi_{d+2}(a_{n+1})) \\ &= (1-\alpha_1)\pi_{d+2}(a_1) + \alpha_1\beta_2 \pi_{d+2}(a_2) + \cdots + \alpha_1\beta_{n+1}\pi_{d+2}(a_{n+1}),\end{aligned}$$

and clearly, $1 - \alpha_1 + \alpha_1\beta_2 + \cdots + \alpha_1\beta_{n+1} = 1$ as $\beta_2 + \cdots + \beta_{n+1} = 1$; $1 - \alpha_1 \geq 0$; and $\alpha_1\beta_i \geq 0$, as $0 \leq \alpha_1 \leq 1$ and $\beta_i \geq 0$. This establishes the induction step, and thus all is left is to prove is the case $n = 2$.

(2) The base case $n = 1$ is also clear. As in (1), let us assume for a moment that (2) is proved for $n = 2$ and consider the induction step. The proof is quite similar to that of (1), but this time, we may assume that $\mu_1 \neq 1$, and we write

$$\begin{aligned}&\mu_1 \pi_{d+2}(a_1) + \cdots + \mu_{n+1}\pi_{d+2}(a_{n+1}) \\ &\qquad = \mu_1 \pi_{d+2}(a_1) + (1-\mu_1)\left(\frac{\mu_2}{1-\mu_1}\pi_{d+2}(a_2) \cdots + \frac{\mu_{n+1}}{1-\mu_1}\pi_{d+2}(a_{n+1})\right).\end{aligned}$$

By the induction hypothesis, there are some $\alpha_2, \ldots, \alpha_{n+1}$ with $\alpha_2 + \cdots + \alpha_{n+1} = 1$ and $\alpha_i \geq 0$ such that

$$\pi_{d+2}(\alpha_2 a_2 + \cdots + \alpha_{n+1} a_{n+1}) = \frac{\mu_2}{1-\mu_1}\pi_{d+2}(a_2) + \cdots + \frac{\mu_{n+1}}{1-\mu_1}\pi_{d+2}(a_{n+1}).$$

By the induction hypothesis for $n = 2$, there is some β_1 with $0 \leq \beta_1 \leq 1$, so that

$$\pi_{d+2}((1-\beta_1)a_1 + \beta_1(\alpha_2 a_2 + \cdots + \alpha_{n+1}a_{n+1})) = \mu_1\pi_{d+2}(a_1) + (1-\mu_1)\pi_{d+2}(\alpha_2 a_2 + \cdots + \alpha_{n+1}a_{n+1}),$$

which establishes the induction hypothesis. Therefore, all that remains is to prove (1) and (2) for $n = 2$.

As π_{d+2} is given by

$$\pi_{d+2}(x_1, \ldots, x_{d+2}) = \left(\frac{x_1}{x_{d+2}}, \ldots, \frac{x_{d+1}}{x_{d+2}}\right) \qquad (x_{d+2} \neq 0),$$

it is enough to treat the case when $d = 0$, that is,

$$\pi_2(e, f) = \frac{e}{f}.$$

Let $a = (a_1, b_1)$ and $b = (a_2, b_2)$. To prove (1), we need to show that for any λ, with $0 \leq \lambda \leq 1$,

$$\pi_2((1-\lambda)a + \lambda b) = \mathrm{conv}(\{\pi_2(a), \pi_2(b)\}).$$

But since

$$\pi_2(a) = \frac{a_1}{b_1}, \qquad \pi_2(b) = \frac{a_2}{b_2}$$

$$\pi_2((1-\lambda)a + \lambda b) = \pi_2((1-\lambda)a_1 + \lambda a_2, (1-\lambda)b_1 + \lambda b_2) = \frac{(1-\lambda)a_1 + \lambda a_2}{(1-\lambda)b_1 + \lambda b_2},$$

it is enough to show that for any λ, with $0 \leq \lambda \leq 1$, if $b_1 b_2 > 0$ then

$$\frac{a_1}{b_1} \leq \frac{(1-\lambda)a_1 + \lambda a_2}{(1-\lambda)b_1 + \lambda b_2} \leq \frac{a_2}{b_2} \quad \text{if} \quad \frac{a_1}{b_1} \leq \frac{a_2}{b_2},$$

and

$$\frac{a_2}{b_2} \leq \frac{(1-\lambda)a_1 + \lambda a_2}{(1-\lambda)b_1 + \lambda b_2} \leq \frac{a_1}{b_1} \quad \text{if} \quad \frac{a_2}{b_2} \leq \frac{a_1}{b_1},$$

where, of course, $(1-\lambda)b_1 + \lambda b_2 \neq 0$. For this, we compute (leaving some steps as an exercise)

$$\frac{(1-\lambda)a_1 + \lambda a_2}{(1-\lambda)b_1 + \lambda b_2} - \frac{a_1}{b_1} = \frac{\lambda(a_2 b_1 - a_1 b_2)}{((1-\lambda)b_1 + \lambda b_2)b_1}$$

and

$$\frac{(1-\lambda)a_1 + \lambda a_2}{(1-\lambda)b_1 + \lambda b_2} - \frac{a_2}{b_2} = -\frac{(1-\lambda)(a_2 b_1 - a_1 b_2)}{((1-\lambda)b_1 + \lambda b_2)b_2}.$$

Now, as $b_1 b_2 > 0$, that is, b_1 and b_2 have the same sign, and as $0 \leq \lambda \leq 1$, we have both $((1-\lambda)b_1 + \lambda b_2)b_1 > 0$ and $((1-\lambda)b_1 + \lambda b_2)b_2 > 0$. Then, if $a_2 b_1 - a_1 b_2 \geq 0$, that is $\frac{a_1}{b_1} \leq \frac{a_2}{b_2}$ (since $b_1 b_2 > 0$), the first two inequalities hold, and if $a_2 b_1 - a_1 b_2 \leq 0$, that is $\frac{a_2}{b_2} \leq \frac{a_1}{b_1}$ (since $b_1 b_2 > 0$), the last two inequalities hold. This proves (1).

In order to prove (2), once again set $a = (a_1, b_1)$ and $b = (a_2, b_2)$. Then given any μ, with $0 \leq \mu \leq 1$, if $b_1 b_2 > 0$, we show that we can find λ with $0 \leq \lambda \leq 1$, so that

$$(1-\mu)\frac{a_1}{b_1} + \mu\frac{a_2}{b_2} = \frac{(1-\lambda)a_1 + \lambda a_2}{(1-\lambda)b_1 + \lambda b_2}$$

If we let

$$\alpha = (1-\mu)\frac{a_1}{b_1} + \mu\frac{a_2}{b_2},$$

we find that λ is given by the equation

$$\lambda(a_2 - a_1 + \alpha(b_1 - b_2)) = \alpha b_1 - a_1.$$

After some (tedious) computations (check for yourself!) we find

$$\begin{aligned} a_2 - a_1 + \alpha(b_1 - b_2) &= \frac{((1-\mu)b_2 + \mu b_1)(a_2 b_1 - a_1 b_2)}{b_1 b_2} \\ \alpha b_1 - a_1 &= \frac{\mu b_1 (a_2 b_1 - a_1 b_2)}{b_1 b_2}. \end{aligned}$$

If $a_2 b_1 - a_1 b_2 = 0$, then $\frac{a_1}{b_1} = \frac{a_2}{b_2}$ and $\lambda = 0$ works. If $a_2 b_1 - a_1 b_2 \neq 0$, then

$$\lambda = \frac{\mu b_1}{(1-\mu)b_2 + \mu b_1} = \frac{\mu}{(1-\mu)\frac{b_2}{b_1} + \mu}.$$

Since $b_1 b_2 > 0$, we have $\frac{b_2}{b_1} > 0$, and since $0 \leq \mu \leq 1$, we conclude that $0 \leq \lambda \leq 1$, which proves (2).

(3) This proof is completely analogous to the proof of (2).

(4) Since

$$\Theta = \pi_{d+2} \circ \widehat{\theta} \circ i_{d+2},$$

as i_{d+2} and $\widehat{\theta}$ are linear, they preserve convex hulls, so by (2), we simply have to show that either $\widehat{\theta} \circ i_{d+2}(\{a_1, \ldots, a_n\})$ is strictly below the hyperplane, $x_{d+2} = 0$, or strictly above it. But

$$\widehat{\theta}(x_1, \ldots, x_{d+2})_{d+2} = x_{d+2} - x_{d+1}$$

and $i_{d+2}(x_1, \ldots, x_{d+1}) = (x_1, \ldots, x_{d+1}, 1)$, so

$$(\widehat{\theta} \circ i_{d+2})(x_1, \ldots, x_{d+1})_{d+2} = 1 - x_{d+1},$$

and this quantity is positive iff $x_{d+1} < 1$, negative iff $x_{d+1} > 1$; that is, either all the points a_i are strictly below the hyperplane H_{d+1} or all strictly above it.

(5) This follows immediately from (4) as $\mathrm{conv}(S)$ consists of all finite convex combinations of points in S. □

If a set $\{a_1, \ldots, a_n\} \subseteq \mathbb{E}^{d+2}$ contains points on *both sides* of the hyperplane $x_{d+2} = 0$, then $\pi_{d+2}(\mathrm{conv}(\{a_1, \ldots, a_n\}))$ is **not** necessarily convex; see Figure 13.24.

Besides θ, we need to define a few more maps in order to establish the connection between the Delaunay complex on S^d and the Delaunay complex on $\mathcal{P}$. We use the convention of denoting the extension to projective spaces of a map f defined between Euclidean spaces by $\widetilde{f}$.

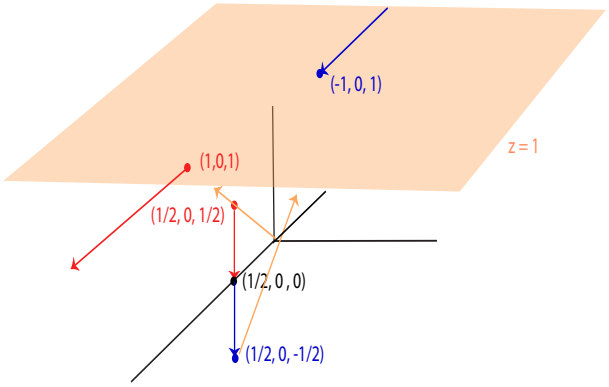


Figure 13.24: Let $a_1 = (1/2, 0, 1/2)$ and $a_2 = (1/2, 0, -1/2)$. Since $\pi_3((1/2, 0, 0))$ is undefined, the image of $\pi_3(\mathrm{conv}(\{a_1, a_2\})$ is two disconnected infinite rays.

Definition 13.7. The Euclidean orthogonal projection $p_i \colon \mathbb{R}^{d+1} \to \mathbb{R}^d$ is given by

$$p_i(x_1, \ldots, x_{d+1}) = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{d+1}),$$

and $\widetilde{p}_i \colon \mathbb{P}^{d+1} \to \mathbb{P}^d$ denotes the projection from $\mathbb{P}^{d+1}$ onto $\mathbb{P}^d$ given by

$$\widetilde{p}_i(x_1 \colon \cdots \colon x_{d+2}) = (x_1 \colon \cdots \colon x_{i-1} \colon x_{i+1} \colon \cdots \colon x_{d+2}),$$

which is undefined at the point $(0 \colon \cdots \colon 1 \colon 0 \colon \cdots \colon 0)$, where the "1" is in the i^{th} slot. The map $\widetilde{\pi}_N \colon (\mathbb{P}^{d+1} - \{\widetilde{N}\}) \to \mathbb{P}^d$ is the central projection from the projective north pole onto $\mathbb{P}^d$ given by

$$\widetilde{\pi}_N(x_1 \colon \cdots \colon x_{d+1} \colon x_{d+2}) = (x_1 \colon \cdots \colon x_d \colon x_{d+2} - x_{d+1}).$$

A geometric interpretation of $\widetilde{\pi}_N$ will be needed later in certain proofs. If we identify $\mathbb{P}^d$ with the hyperplane $H_{d+1}(0) \subseteq \mathbb{P}^{d+1}$ of equation $x_{d+1} = 0$, then we claim that for any $x \neq \widetilde{N}$, the point $\widetilde{\pi}_N(x)$ is the intersection of the line through $\widetilde{N}$ and x with the hyperplane $H_{d+1}(0)$. See Figure 13.25. Indeed, parametrically, the line $\langle \widetilde{N}, x \rangle$ through $\widetilde{N} = (0 \colon \cdots \colon 0 \colon 1 \colon 1)$ and x is given by

$$\langle \widetilde{N}, x \rangle = \{(\mu x_1 \colon \cdots \colon \mu x_d \colon \lambda + \mu x_{d+1} \colon \lambda + \mu x_{d+2}) \mid \lambda, \mu \in \mathbb{R},\ \lambda \neq 0 \quad \text{or} \quad \mu \neq 0\}.$$

The line $\langle \widetilde{N}, x \rangle$ intersects the hyperplane $x_{d+1} = 0$ iff

$$\lambda + \mu x_{d+1} = 0,$$

so we can pick $\lambda = -x_{d+1}$ and $\mu = 1$, which yields the intersection point,

$$(x_1 \colon \cdots \colon x_d \colon 0 \colon x_{d+2} - x_{d+1}),$$

as claimed.

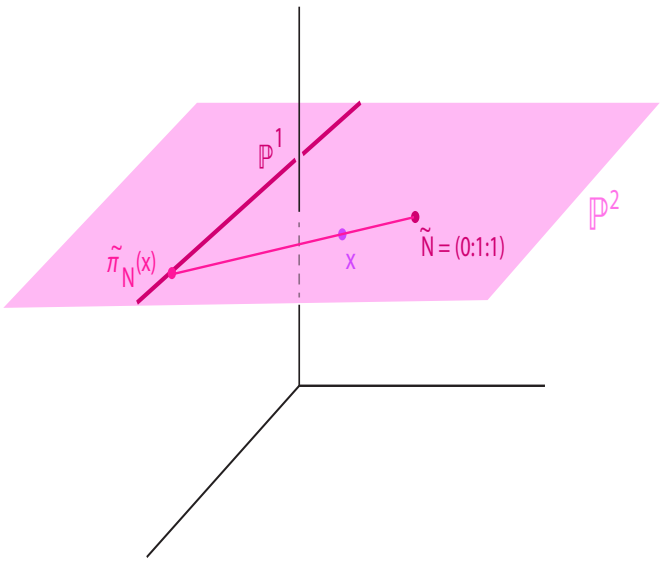


Figure 13.25: A schematic representation, via the plane model for $\mathbb{P}^2$, of the geometric image of $\widetilde{\pi}_N(x)$. The copy of $\mathbb{P}^1$ corresponds to the intersection of the xz- plane with the plane $z = 1$.

Definition 13.8. The projective versions of σ_N and τ_N, denoted $\widetilde{\sigma}_N \colon (\widetilde{S^d} - \{\widetilde{N}\}) \to \mathbb{P}^d$ and $\widetilde{\tau}_N \colon \mathbb{P}^d \to \widetilde{S^d} \subseteq \mathbb{P}^{d+1}$, are given by

$$\widetilde{\sigma}_N(x_1 \colon \cdots \colon x_{d+2}) = (x_1 \colon \cdots \colon x_d \colon x_{d+2} - x_{d+1}),$$

and

$$\widetilde{\tau}_N(x_1 \colon \cdots \colon x_{d+1}) = \left(2x_1x_{d+1} \colon \cdots \colon 2x_dx_{d+1} \colon \sum_{i=1}^{d} x_i^2 - x_{d+1}^2 \colon \sum_{i=1}^{d} x_i^2 + x_{d+1}^2\right).$$

It is an easy exercise to check that the image of $\widetilde{S^d} - \{\widetilde{N}\}$ by $\widetilde{\sigma}_N$ is U_{d+1}, and that $\widetilde{\sigma}_N$ and $\widetilde{\tau}_N \restriction U_{d+1}$ are mutual inverses.

Observe that $\widetilde{\sigma}_N = \widetilde{\pi}_N \restriction \widetilde{S^d}$, the restriction of the projection $\widetilde{\pi}_N$ to the sphere $\widetilde{S^d}$.

Definition 13.9. The lifting $\widetilde{l} \colon \mathbb{E}^d \to \widetilde{\mathcal{P}} \subseteq \mathbb{P}^{d+1}$ is given by

$$\widetilde{l}(x_1, \ldots, x_d) = \left(x_1 \colon \cdots \colon x_d \colon \sum_{i=1}^{d} x_i^2 \colon 1\right),$$

and the embedding $\psi_{d+1} \colon \mathbb{E}^d \to \mathbb{P}^d$ (the map ψ_{d+1} defined in Section 12.1) is given by

$$\psi_{d+1}(x_1, \ldots, x_d) = (x_1 \colon \cdots \colon x_d \colon 1).$$

Then, we easily check the following facts.

Proposition 13.6. *The maps $\theta, \widetilde{\pi}_N, \widetilde{\tau}_N, \widetilde{p}_{d+1}, \widetilde{l}$ and ψ_{d+1} defined before satisfy the equations*

$$\begin{aligned}
\widetilde{l} &= \theta \circ \widetilde{\tau}_N \circ \psi_{d+1} \\
\widetilde{\pi}_N &= \widetilde{p}_{d+1} \circ \theta \\
\widetilde{\tau}_N \circ \psi_{d+1} &= \psi_{d+2} \circ \tau_N \\
\widetilde{l} &= \psi_{d+2} \circ l \\
l &= \Theta \circ \tau_N.
\end{aligned}$$

Proof. Recall that θ is given by

$$\theta(x_1 \colon \cdots \colon x_{d+2}) = (x_1 \colon \cdots \colon x_d \colon x_{d+1} + x_{d+2} \colon x_{d+2} - x_{d+1}).$$

Then, as

$$\widetilde{\tau}_N \circ \psi_{d+1}(x_1, \ldots, x_d) = \left(2x_1 \colon \cdots \colon 2x_d \colon \sum_{i=1}^{d} x_i^2 - 1 \colon \sum_{i=1}^{d} x_i^2 + 1\right),$$

we get

$$\begin{aligned}
\theta \circ \widetilde{\tau}_N \circ \psi_{d+1}(x_1, \ldots, x_d) &= \left(2x_1 \colon \cdots \colon 2x_d \colon 2\sum_{i=1}^{d} x_i^2 \colon 2\right) \\
&= \left(x_1 \colon \cdots \colon x_d \colon \sum_{i=1}^{d} x_i^2 \colon 1\right) = \widetilde{l}(x_1, \ldots, x_d),
\end{aligned}$$

which proves the first equation.

For the second equation, since $\widetilde{p}_{d+1}$ drops the $(d+1)$th component of a $(d+2)$-tuple, the equation $\widetilde{\pi}_N = \widetilde{p}_{d+1} \circ \theta$ follows immediately from the definitions.

We have

$$\begin{aligned}
\widetilde{\tau}_N(\psi_{d+1}(x)) &= \widetilde{\tau}_N(x_1 \colon \cdots \colon x_d \colon 1) \\
&= (2x_1 \colon \cdots \colon 2x_d \colon \|x\|^2 - 1 \colon \|x\|^2 + 1),
\end{aligned}$$

and

$$\begin{aligned}
\psi_{d+2}(\tau_N(x)) &= \left(\frac{2x_1}{\|x\|^2 + 1} \colon \cdots \colon \frac{2x_d}{\|x\|^2 + 1} \colon \frac{\|x\|^2 - 1}{\|x\|^2 + 1} \colon 1\right) \\
&= (2x_1 \colon \cdots \colon 2x_d \colon \|x\|^2 - 1 \colon \|x\|^2 + 1),
\end{aligned}$$

so the third equation holds.

Since ψ_{d+2} adds a 1 as a $(d+2)$th component, the fourth equation follows immediately from the definitions.

For the fifth equation, since

$$\tau_N(x) = \left(\frac{2x}{\|x\|^2 + 1}, \frac{\|x\|^2 - 1}{\|x\|^2 + 1} \right)$$

and

$$\begin{aligned} \Theta(x)_i &= \frac{x_i}{1 - x_{d+1}}, \qquad 1 \leq i \leq d \\ \Theta(x)_{d+1} &= \frac{x_{d+1} + 1}{1 - x_{d+1}}, \end{aligned}$$

we get

$$\begin{aligned} \Theta(\tau_N(x))_i &= \left(\frac{2x_i}{\|x\|^2 + 1} \right) \Big/ \left(1 - \frac{\|x\|^2 - 1}{\|x\|^2 + 1} \right) \\ &= \frac{2x_i}{2} = x_i, \quad 1 \leq i \leq d, \end{aligned}$$

$$\begin{aligned} \theta(\tau_N(x))_{d+1} &= \left(\frac{\|x\|^2 - 1}{\|x\|^2 + 1} + 1 \right) \Big/ \left(1 - \frac{\|x\|^2 - 1}{\|x\|^2 + 1} \right) \\ &= \frac{2\|x\|^2}{2} = \|x\|^2, \end{aligned}$$

and since

$$l(x) = (x_1, \ldots, x_d, \sum_{i=1}^{d} x_i^2) = (x_1, \ldots, x_d, \|x\|^2),$$

we have shown that $l = \Theta \circ \tau_N$, as claimed. □

13.7 Lifted Delaunay Complexes and Delaunay Complexes via Lifting to a Sphere

In order to define precisely Delaunay complexes as projections of objects obtained by deleting some faces from a projective polyhedron we need to define the notion of "projective (polyhedral) complex." However, this is easily done by defining the notion of cell complex where the cells are polyhedral cones. Such objects are known as *fans*. The definition below is basically Definition 10.10 in which the cells are cones as opposed to polytopes.

Definition 13.10. A *fan in* $\mathbb{E}^m$ is a set K consisting of a (finite or infinite) set of polyhedral cones in $\mathbb{E}^m$ satisfying the following conditions:

(1) Every face of a cone in K also belongs to K.

(2) For any two cones 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. See Figure 13.26.

Every cone $\sigma \in K$ 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 fan K is the maximum of the dimensions of all cones 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*.

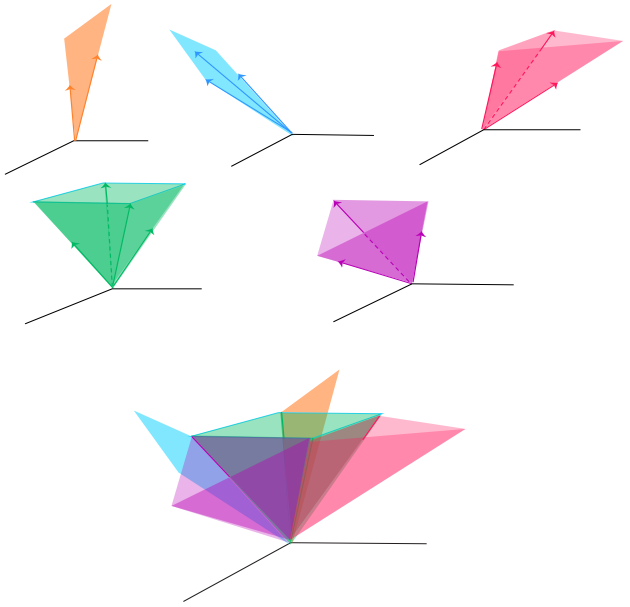


Figure 13.26: Five polyhedral cones in $\mathbb{E}^3$ combined into a fan.

A *projective (polyhedral) complex* $\mathcal{K} \subseteq 2^{\mathbb{P}^d}$ is a set of projective polyhedra of the form $\{\mathbb{P}(C) \mid C \in K\}$, where $K \subseteq 2^{\mathbb{R}^{d+1}}$ is a fan.

Given a projective complex, the notions of face, vertex, edge, cell, facet, are defined in the obvious way.

If $K \subseteq 2^{\mathbb{R}^d}$ is a polyhedral complex, then it is easy to check that the set $\{C(\sigma) \mid \sigma \in K\} \subseteq 2^{\mathbb{R}^{d+1}}$ (where $C(\sigma)$ is the $\mathcal{V}$-cone associated with σ defined in Section 5.5) is a fan.

Definition 13.11. Given a polyhedral complex $K \subseteq 2^{\mathbb{R}^d}$, the projective complex

$$\widetilde{K} = \{\mathbb{P}(C(\sigma)) \mid \sigma \in K\} \subseteq 2^{\mathbb{P}^d}$$

is called the *projective completion* of K. See Figure 13.27.

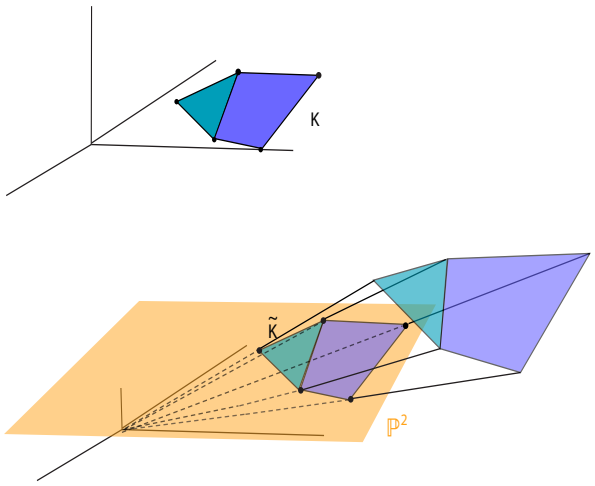


Figure 13.27: The projective completion, in the plane model of $\mathbb{P}^2$, of K, the two dimensional complex in the xy-plane consisting of the teal triangle and the periwinkle quadrilateral.

Also, it is easy to check that if $f\colon P \to P'$ is an injective affine map between two polyhedra P and P', then f extends uniquely to a projectivity $\widetilde{f}\colon \widetilde{P} \to \widetilde{P'}$ between the projective completions of P and P'.

We now have all the facts needed to show that Delaunay triangulations and Voronoi diagrams can be defined in terms of the lifting $\widetilde{\tau}_N \circ \psi_{d+1}$, and the projection $\widetilde{\pi}_N$, and to establish their duality *via* polar duality with respect to S^d.

Definition 13.12. Given any set of points $P = \{p_1, \ldots, p_n\} \subseteq \mathbb{E}^d$, the polytope $\mathcal{D}(P) \subseteq \mathbb{R}^{d+1}$, called the *Delaunay polytope* associated with P, is the convex hull of the union of the lifting of the points of P onto the sphere S^d (*via* inverse stereographic projection) with the north pole; that is, $\mathcal{D}(P) = \mathrm{conv}(\tau_N(P) \cup \{N\})$. See Figures 13.28, 13.29, and 13.30. The *projective Delaunay polytope* $\widetilde{\mathcal{D}}(P) \subseteq \mathbb{P}^{d+1}$ associated with P is the projective completion of $\mathcal{D}(P)$. The polyhedral complex $\mathcal{DC}(P) \subseteq 2^{\mathbb{R}^{d+1}}$, called the *lifted Delaunay complex of* P, is the complex obtained from the boundary of $\mathcal{D}(P)$ by deleting the facets containing the north pole (and their faces), as illustrated in Figure 13.31, and $\widetilde{\mathcal{DC}}(P) \subseteq 2^{\mathbb{P}^{d+1}}$ is the projective completion of $\mathcal{DC}(P)$. The polyhedral complex $\mathcal{D}el(P) = \varphi_{d+1} \circ \widetilde{\pi}_N(\widetilde{\mathcal{DC}}(P)) \subseteq 2^{\mathbb{E}^d}$ is the *Delaunay complex of* P.

The above is not the "standard" definition of the Delaunay triangulation of P, but it is equivalent to the definition given in Section 17.3.1 of Boissonnat and Yvinec [12], as we will prove shortly. It also has certain advantages over lifting onto a paraboloid, as we will explain. Furthermore, to be perfectly rigorous, we should define $\mathcal{D}el(P)$ by

$$\mathcal{D}el(P) = \varphi_{d+1}(\widetilde{\pi}_N(\widetilde{\mathcal{DC}}(P)) \cap 2^{U_{d+1}}),$$

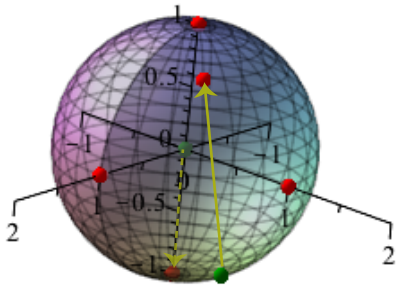


Figure 13.28: The inverse stereographic projection of $P = \{(0,0,0),(1,0,0),(0,1,0),(2,2,0)\}$. Note that $\tau_N((0,0,0)) = (0,0,-1)$ and $\tau_N((2,2,0)) = (4/9,4/9,7/9)$.

but $\widetilde{\pi}_N(\widetilde{\mathcal{DC}}(P)) \subseteq 2^{U_{d+1}}$ because $\mathcal{DC}(P)$ is strictly below the hyperplane H_{d+1}.

The Delaunay complex $\mathcal{D}el(P)$ is often called the *Delaunay triangulation of* P, but this terminology is slightly misleading because $\mathcal{D}el(P)$ is not simplicial unless P is in general position; see Proposition 13.11.

It it possible and useful to define $\mathcal{D}el(P)$ more directly in terms of $\mathcal{DC}(P)$. The projection $\widetilde{\pi_N}\colon (\mathbb{P}^{d+1} - \{\widetilde{N}\}) \to \mathbb{P}^d$ comes from the linear map $\widehat{\pi}_N\colon \mathbb{R}^{d+2} \to \mathbb{R}^{d+1}$ given by

$$\widehat{\pi}_N(x_1,\ldots,x_{d+1},x_{d+2}) = (x_1,\ldots,x_d,x_{d+2}-x_{d+1}).$$

Consequently, as $\widetilde{\mathcal{DC}}(P) = \widetilde{\mathcal{DC}(P)} = \mathbb{P}(C(\mathcal{DC}(P)))$, we immediately check that

$$\mathcal{D}el(P) = \varphi_{d+1} \circ \widetilde{\pi}_N(\widetilde{\mathcal{DC}}(P)) = \varphi_{d+1} \circ \widehat{\pi}_N(C(\mathcal{DC}(P))) = \varphi_{d+1} \circ \widehat{\pi}_N(\mathrm{cone}(\widehat{\mathcal{DC}(P)})),$$

where $\widehat{\mathcal{DC}(P)} = \{\widehat{u} \mid u \in \mathcal{DC}(P)\}$ and $\widehat{u} = (u,1)$.

This suggests defining the map $\pi_N\colon (\mathbb{R}^{d+1} - H_{d+1}) \to \mathbb{R}^d$ by

$$\pi_N = \varphi_{d+1} \circ \widehat{\pi}_N \circ i_{d+2},$$

which is explicity given by

$$\pi_N(x_1,\ldots,x_d,x_{d+1}) = \frac{1}{1-x_{d+1}}(x_1,\ldots,x_d).$$

Observe that the map π_N is just the central projection from the north pole to the hyperplane $x_{d+1} = 0$. Then, as $\mathcal{DC}(P)$ is strictly below the hyperplane H_{d+1}, we have

$$\mathcal{D}el(P) = \varphi_{d+1} \circ \widetilde{\pi}_N(\widetilde{\mathcal{DC}}(P)) = \pi_N(\mathcal{DC}(P)).$$

See Figures 13.31 and 13.32.

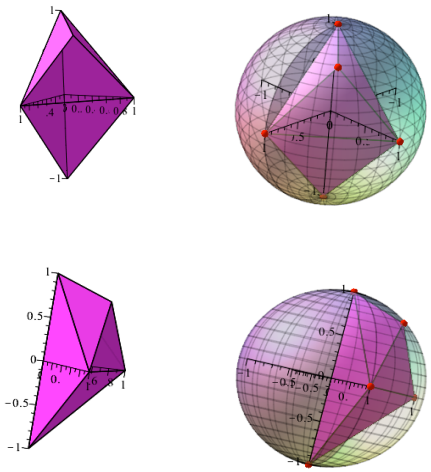


Figure 13.29: Two views of the Delaunay polytope $\mathcal{D}(P)$, where $P = \{(0,0,0),(1,0,0),(0,1,0),(2,2,0)\}$.

First, note that $\mathcal{D}el(P) = \varphi_{d+1} \circ \widetilde{\pi}_N(\widetilde{\mathcal{DC}}(P)) = \pi_N(\mathcal{DC}(P))$ is indeed a polyhedral complex whose geometric realization is the convex hull $\mathrm{conv}(P)$ of P. Indeed, by Proposition 13.5, the images of the facets of $\mathcal{DC}(P)$ are polytopes, and when any two such polytopes meet, they meet along a common face. Furthermore, if $\dim(\mathrm{conv}(P)) = m$, then $\mathcal{D}el(P)$ is pure m-dimensional. First, $\mathcal{D}el(P)$ contains at least one m-dimensional cell. If $\mathcal{D}el(P)$ was not pure, as the complex is connected there would be some cell σ of dimension $s < m$ meeting some other cell τ of dimension m along a common face of dimension at most s, and because σ is not contained in any face of dimension m, no facet of τ containing $\sigma \cap \tau$ can be adjacent to any cell of dimension m, and so $\mathcal{D}el(P)$ would not be convex, a contradiction.

Our next goal is to show that the Delaunay complex $\mathcal{D}el(P)$ coincides with the "standard" Delaunay complex $\mathcal{D}el'(P)$, as defined in Section 17.3.1 of Boissonnat and Yvinec [12]. To define $\mathcal{D}el'(P)$, we need the notion of lower-facing facet.

For any polytope $P \subseteq \mathbb{E}^d$, given any point x not in P, recall that a facet F of P is *visible from* x iff for every point $y \in F$, the line through x and y intersects P only in y. If $\dim(P) = d$, this is equivalent to saying that x and the interior of P are strictly separated by the supporting hyperplane of F. Note that if $\dim(P) < d$, it possible that every facet of P is visible from x. See Figure 13.33.

Definition 13.13. Assume that $P \subseteq \mathbb{E}^d$ is a polytope with $\dim(P) = d$. We say that a facet F of P is a *lower-facing facet* of P iff the unit normal to the supporting hyperplane of F pointing towards the interior of P has non-negative x_{d+1}-coordinate. A facet F that is not lower-facing is called an *upper-facing facet* (note that in this case, the x_{d+1} coordinate of the unit normal to the supporting hyperplane of F pointing towards the interior of P is strictly negative). See Figure 13.34.

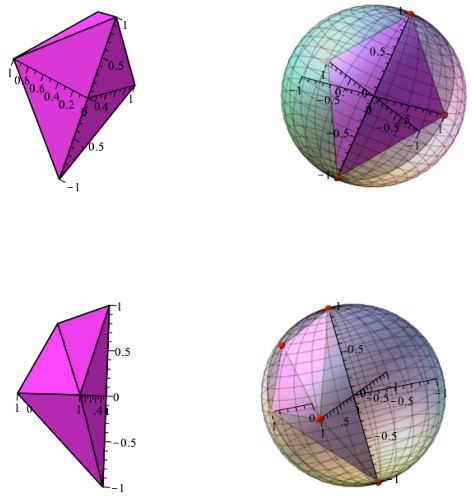


Figure 13.30: Two additional views of the Delaunay polytope $\mathcal{D}(P)$ circumscribed by the unit sphere.

Here is a convenient way to characterize lower-facing facets.

Proposition 13.7. *Given any polytope $P \subseteq \mathbb{E}^d$ with $\dim(P) = d$, for any point c on the Ox_d-axis, if c lies strictly above all the intersection points of the Ox_d-axis with the supporting hyperplanes of all the upper-facing facets of F, then the lower-facing facets of P are exactly the facets not visible from c. See Figure 13.35.*

Proof. Note that the intersection points of the Ox_d-axis with the supporting hyperplanes of all the upper-facing facets of P are strictly above the intersection points of the Ox_d-axis with the supporting hyperplanes of all the lower-facing facets. Suppose F is visible from c. Then, F must not be lower-facing, as otherwise, for any $y \in F$, the line through c and y has to intersect some upper-facing facet and F is not be visible from c, a contradiction.

Now, as P is the intersection of the closed half-spaces determined by the supporting hyperplanes of its facets, by the definition of an upper-facing facet, any point c on the Ox_d-axis that lies strictly above the intersection points of the Ox_d-axis with the supporting hyperplanes of all the upper-facing facets of F has the property that c and the interior of P are strictly separated by all these supporting hyperplanes. Therefore, all the upper-facing facets of P are visible from c. It follows that the facets visible from c are exactly the upper-facing facets, as claimed. □

We will also need the following fact.

Proposition 13.8. *Given any polytope $P \subseteq \mathbb{E}^d$, if $\dim(P) = d$, then there is a point c on the Ox_d-axis such that for all points x on the Ox_d-axis and above c, the set of facets of*

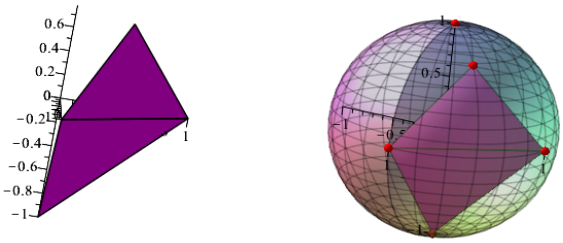


Figure 13.31: The lifted Delaunay complex $\mathcal{DC}(P)$, where $P = \{(0,0,0),(1,0,0),(0,1,0),(2,2,0)\}$.

$\mathrm{conv}(P \cup \{x\})$ *not containing* x *is identical. Moreover, the set of facets of* P *not visible from* x *is the set of facets of* $\mathrm{conv}(P \cup \{x\})$ *that do not contain* x.

Proof. If $\dim(P) = d$ then pick any c on the Ox_d-axis above the intersection points of the Ox_d-axis with the supporting hyperplanes of all the upper-facing facets of F. Then, c is in general position w.r.t. P in the sense that c and any d vertices of P do not lie in a common hyperplane. Now, our result follows by Lemma 8.3.1 of Boissonnat and Yvinec [12]. □

Corollary 13.9. *Given any polytope* $P \subseteq \mathbb{E}^d$ *with* $\dim(P) = d$*, there is a point* c *on the* Ox_d*-axis so that for all* x *on the* Ox_d*-axis and above* c*, the lower-facing facets of* P *are exactly the facets of* $\mathrm{conv}(P \cup \{x\})$ *that do not contain* x*. See Figure 13.36.*

As usual, let $e_{d+1} = (0,\ldots,0,1) \in \mathbb{R}^{d+1}$.

The standard Delaunay polyhedron and the standard Delaunay complex are defined as follows (compare Boissonnat and Yvinec [12], Section 17.3.1).

Definition 13.14. Given any set of points $P = \{p_1,\ldots,p_n\} \subseteq \mathbb{E}^d$, let $\mathcal{D}'(P)$ denote the polyhedron $\mathrm{conv}(l(P)) + \mathrm{cone}(e_{d+1})$, and let $\widetilde{\mathcal{D}}'(P)$ be the projective completion of $\mathcal{D}'(P)$. Also, let $\mathcal{DC}'(P)$ be the polyhedral complex consisting of the bounded facets of the polytope $\mathcal{D}'(P)$, and let $\widetilde{\mathcal{DC}}'(P)$ be the projective completion of $\mathcal{DC}'(P)$. See Figure 13.37. The complex $\mathcal{D}el'(P) = \varphi_{d+1} \circ \widetilde{p}_{d+1}(\widetilde{\mathcal{DC}}'(P)) = p_{d+1}(\mathcal{DC}'(P))$ is the *standard Delaunay complex* of P, that is, the orthogonal projection of $\mathcal{DC}'(P)$ onto $\mathbb{E}^d$. See Figure 13.38.

Intuitively, adding to $\mathrm{conv}(l(P))$ all the verttical rays parallel to e_{d+1} based on points in $\mathrm{conv}(l(P))$ washes out the upper-facing faces of $\mathrm{conv}(l(P))$. Then the bounded facets of $\mathrm{conv}(l(P))+\mathrm{cone}(e_{d+1})$ are precisely the lower-facing facets of $\mathrm{conv}(l(P))$ (if $\dim(\mathrm{conv}(P)) = d$).

The first of the two main theorems of this chapter is that the two notions of Delaunay complexes coincide.

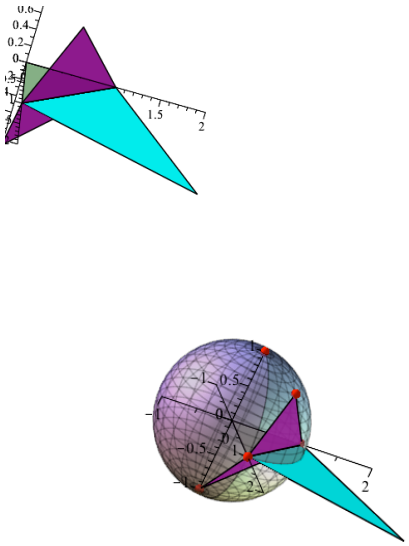


Figure 13.32: The Delaunay complex $\mathcal{D}el(P)$, for $P = \{(0,0,0),(1,0,0),(0,1,0),(2,2,0)\}$ obtained by applying π_N to $\mathcal{DC}(P)$ of Figure 13.31. The bottom triangle projects onto the planar green triangle with vertices $\{(0,0,0),(1,0,0),(0,1,0)\}$, while the top triangle projection onto the aqua triangle with vertices $\{(1,0,0),(0,1,0),(2,2,0)\}$.

Theorem 13.10. *Given any set of points $P = \{p_1,\ldots,p_n\} \subseteq \mathbb{E}^d$, we have*

$$\theta(\widetilde{\mathcal{D}}(P)) = \widetilde{\mathcal{D}}'(P) \quad \textit{and} \quad \theta(\widetilde{\mathcal{DC}}(P)) = \widetilde{\mathcal{DC}}'(P).$$

Furthermore,

$$\mathcal{D}el(P) = \mathcal{D}el'(P).$$

Therefore, the two notions of a Delaunay complex agree. If $\dim(\mathrm{conv}(P)) = d$*, then the bounded facets of* $\mathrm{conv}(l(P)) + \mathrm{cone}(e_{d+1})$ *are precisely the lower-facing facets of* $\mathrm{conv}(l(P))$*.*

Proof. Recall that

$$\mathcal{D}(P) = \mathrm{conv}(\tau_N(P) \cup \{N\}),$$

and $\widetilde{\mathcal{D}}(P) = \mathbb{P}(C(\mathcal{D}(P)))$ is the projective completion of $\mathcal{D}(P)$. If we write $\widehat{\tau_N(P)}$ for $\{\widehat{\tau_N(p_i)} \mid p_i \in P\}$, then

$$C(\mathcal{D}(P)) = \mathrm{cone}(\widehat{\tau_N(P)} \cup \{\widehat{N}\}).$$

By definition, we have

$$\theta(\widetilde{\mathcal{D}}) = \mathbb{P}(\widehat{\theta}(C(\mathcal{D}))).$$

Now, as $\widehat{\theta}$ is linear,

$$\widehat{\theta}(C(\mathcal{D})) = \widehat{\theta}(\mathrm{cone}(\widehat{\tau_N(P)} \cup \{\widehat{N}\})) = \mathrm{cone}(\widehat{\theta}(\widehat{\tau_N(P)}) \cup \{\widehat{\theta}(\widehat{N})\}).$$

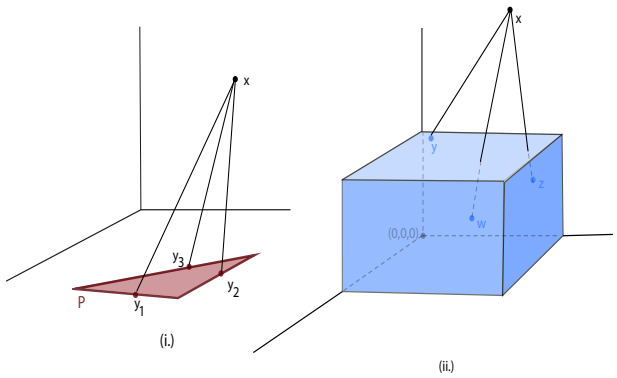


Figure 13.33: In $\mathbb{E}^3$, all three edges of the planar triangle in Figure (i.) are visible from x, while in Figure (ii.), only the top face of the solid rectangular box is visible from x.

We claim that

$$\begin{aligned}\mathrm{cone}(\widehat{\theta}(\widehat{\tau_N(P)}) \cup \{\widehat{\theta}(\widehat{N})\}) &= \mathrm{cone}(\widehat{l(P)} \cup \{(0,\ldots,0,1,1)\}) \\ &= C(\mathcal{D}'(P)),\end{aligned}$$

where

$$\mathcal{D}'(P) = \mathrm{conv}(l(P)) + \mathrm{cone}(e_{d+1}).$$

Indeed,

$$\widehat{\theta}(x_1,\ldots,x_{d+2}) = (x_1,\ldots,x_d, x_{d+1}+x_{d+2}, x_{d+2}-x_{d+1}),$$

and for any $p_i = (x_1,\ldots,x_d) \in P$,

$$\begin{aligned}\widehat{\tau_N(p_i)} &= \left(\frac{2x_1}{\sum_{i=1}^d x_i^2+1},\ldots,\frac{2x_d}{\sum_{i=1}^d x_i^2+1},\frac{\sum_{i=1}^d x_i^2-1}{\sum_{i=1}^d x_i^2+1},1\right) \\ &= \frac{1}{\sum_{i=1}^d x_i^2+1}\left(2x_1,\ldots,2x_d,\sum_{i=1}^d x_i^2-1,\sum_{i=1}^d x_i^2+1\right),\end{aligned}$$

so we get

$$\widehat{\theta}(\widehat{\tau_N(p_i)}) = \frac{2}{\sum_{i=1}^d x_i^2+1}\left(x_1,\ldots,x_d,\sum_{i=1}^d x_i^2,1\right) = \frac{2}{\sum_{i=1}^d x_i^2+1}\,\widehat{l(p_i)}.$$

Also, we have

$$\widehat{\theta}(\widehat{N}) = \widehat{\theta}(0,\ldots,0,1,1) = (0,\ldots,0,2,0) = 2\widehat{e_{d+1}},$$

and by definition of cone$(-)$ (scalar factors are irrelevant), we get

$$\mathrm{cone}(\widehat{\theta}(\widehat{\tau_N(P)}) \cup \{\widehat{\theta}(\widehat{N})\}) = \mathrm{cone}(\widehat{l(P)} \cup \{(0,\ldots,0,1,1)\}) = C(\mathcal{D}'(P)),$$

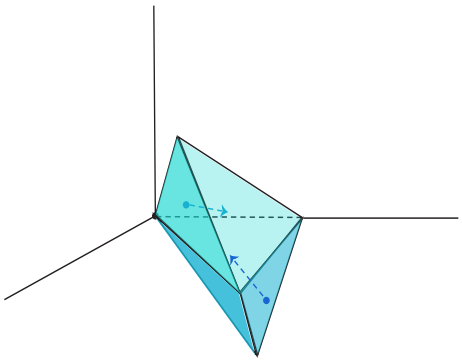


Figure 13.34: Let P be the solid mint green triangular bipyramid. The three faces on the top are upper-facing, while the three faces on the bottom are lower-facing.

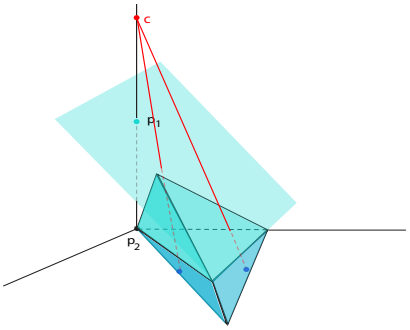


Figure 13.35: Let P be the solid mint green triangular bipyramid. The intersections of the upper-facing facets and the z-axis is given by the two points p_1 and p_2. Since c is above p_1, none of the three lower-facing facets in the bottom of the bipyramid are visible from c.

with $\mathcal{D}'(P) = \mathrm{conv}(l(P)) + \mathrm{cone}(e_{d+1})$, as claimed. This proves that

$$\theta(\widetilde{\mathcal{D}}(P)) = \widetilde{\mathcal{D}}'(P).$$

Now, it is clear that the facets of $\mathrm{conv}(\tau_N(P) \cup \{N\})$ that do not contain N are mapped to the bounded facets of $\mathrm{conv}(l(P)) + \mathrm{cone}(e_{d+1})$, since N goes the point at infinity, so

$$\theta(\widetilde{\mathcal{DC}}(P)) = \widetilde{\mathcal{DC}}'(P).$$

As $\widetilde{\pi}_N = \widetilde{p}_{d+1} \circ \theta$, by Proposition 13.6, we get

$$\mathcal{D}el'(P) = \varphi_{d+1} \circ \widetilde{p}_{d+1}(\widetilde{\mathcal{DC}}'(P)) = \varphi_{d+1} \circ (\widetilde{p}_{d+1} \circ \theta)(\widetilde{\mathcal{DC}}(P)) = \varphi_{d+1} \circ \widetilde{\pi}_N(\widetilde{\mathcal{DC}}(P)) = \mathcal{D}el(P),$$

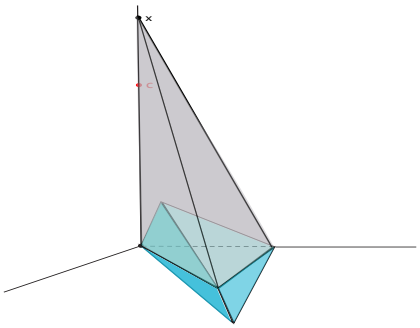


Figure 13.36: Let P be the solid mint green triangular bipyramid. Then $\mathrm{conv}(P \cup \{x\})$ is the larger solid bipyramid with gray top and mint green bottom. The lower-facing facets of P are the three mint green faces on the bottom of both P and $\mathrm{conv}(P \cup \{x\})$.

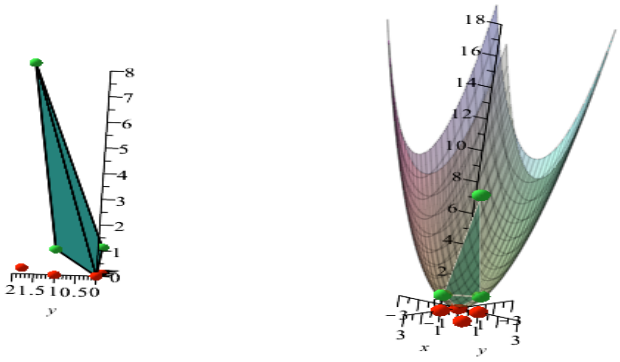


Figure 13.37: Two views of the tetrahedron $l(P)$, where $P = \{(0,0,0),(1,0,0),(0,1,0),(2,2,0)\}$.

as claimed. Finally, if $\dim(\mathrm{conv}(P)) = d$, then, by Corollary 13.9, we can pick a point c on the Ox_{d+1}-axis so that the facets of $\mathrm{conv}(l(P) \cup \{c\})$ that do not contain c are precisely the lower-facing facets of $\mathrm{conv}(l(P))$. However, it is also clear that the facets of $\mathrm{conv}(l(P) \cup \{c\})$ that contain c tend to the unbounded facets of $\mathcal{DC}'(P) = \mathrm{conv}(l(P)) + \mathrm{cone}(e_{d+1})$ when c goes to $+\infty$. □

We can also characterize when the Delaunay complex $\mathcal{D}el(P)$ is simplicial. Recall that we say that a set of points $P \subseteq \mathbb{E}^d$ is in *general position* if no $d+2$ of the points in P belong to a common $(d-1)$-sphere.

Proposition 13.11. *Given any set of points $P = \{p_1, \ldots, p_n\} \subseteq \mathbb{E}^d$, if P is in general position, then the Delaunay complex $\mathcal{D}el(P)$ is a pure simplicial complex. The lifted Delaunay complex $\mathcal{DC}(P)$ is also a pure simplicial complex.*

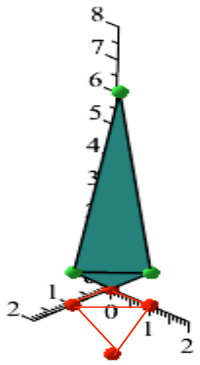


Figure 13.38: The polyhedral complex $\mathcal{DC}'(P)$ for $P = \{(0,0,0),(1,0,0),(0,1,0),(2,2,0)\}$, and its orthogonal projection (in red) onto the xy-plane. Note this orthogonal projection gives the same Delaunay complex as in Figure 13.32.

Proof. Let $\dim(\text{conv}(P)) = r$. Then, $\tau_N(P)$ is contained in a $(r-1)$-sphere of S^d, so we may assume that $r = d$. Suppose $\mathcal{D}el(P)$ has some facet F which is not a d-simplex. If so, F is the convex hull of at least $d+2$ points $p_1, \ldots, p_k$ of P, and since $F = \pi_N(\widehat{F})$ for some facet $\widehat{F}$ of $\mathcal{DC}(P)$, we deduce that $\tau_N(p_1), \ldots, \tau_N(p_k)$ belong to the supporting hyperplane H of $\widehat{F}$. Now, if H passes through the north pole, then we know that $p_1, \ldots, p_k$ belong to some hyperplane of $\mathbb{E}^d$, which is impossible since $p_1, \ldots, p_k$ are the vertices of a facet of dimension d. Thus, H does not pass through N, and so $p_1, \ldots, p_k$ belong to some $(d-1)$-sphere in $\mathbb{E}^d$. As $k \geq d+2$, this contradicts the assumption that the points in P are in general position.

The proof that $\mathcal{DC}(P)$ is a pure simplicial complex is similar. A similar proof is also given in Boissonnat and Yvinec [12], Section 17.3.2. □

Remark: Even when the points in P are in general position, the Delaunay complex $\mathcal{D}(P)$ may not be a simplicial polytope. For example, if $d+1$ points belong to a hyperplane in $\mathbb{E}^d$, then the lifted points belong to a hyperplane passing through the north pole, and these $d+1$ lifted points together with N form a non-simplicial facet. For example, consider the polytope obtained by lifting our original $d+1$ points on a hyperplane H plus one more point not in the hyperplane H; see Figure 13.39.

13.8 Lifted Voronoi Complexes and Voronoi Complexes via Lifting to a Sphere

Our final goal is to characterize the Voronoi diagram of P in terms of the polar dual $\mathcal{D}(P)^*$ of $\mathcal{D}(P)$ (with respect to the sphere S^d). The polar dual $\mathcal{D}(P)^*$ of $\mathcal{D}(P)$ is the polyhedron obtained by intersecting the half-spaces containing the origin associated with the tangent

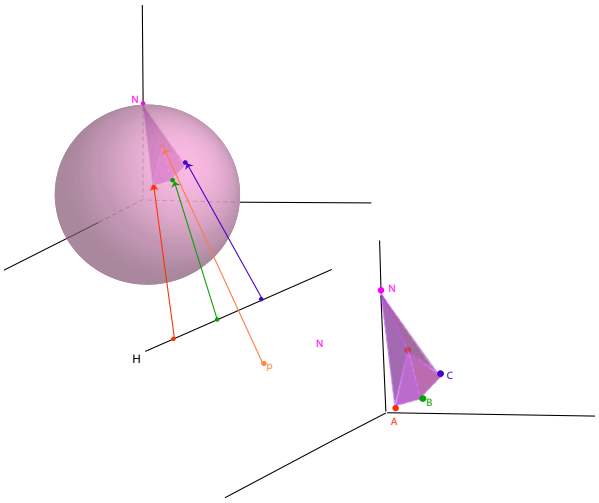


Figure 13.39: Let $d = 2$. The three points on the line H and the orange point p are in general position. However, when lifted to S^2, the convex hull of these three points and N form a solid pyramid with base $ABCN$.

hyperplanes to S^d at the lifted points $\tau_N(p_i)$ (with $p_i \in P$), and at the north pole N. See Figures 13.40, and 13.41. It follows that the polyhedron $\mathcal{D}(P)^*$ has exactly one facet containing the north pole. The Voronoi diagram of P is the result of applying the central projection π_N from N to the polyhedron $\mathcal{D}(P)^*$. Under this central projection, the facet containing the north pole goes to infinity, so instead of considering the polar dual $\mathcal{D}(P)^*$ we should consider the polar dual $\mathcal{DC}(P)^*$ of the lifted Delaunay complex $\mathcal{DC}(P)$ which does not have the north pole as a vertex. Then the Voronoi diagram of P is the result of applying the central projection π_N from N to the complex $\mathcal{DC}(P)^*$. See Figures 13.42 through 13.45.

The polyhedron $\mathcal{DC}(P)^*$ still contains faces intersecting the tangent hyperplane to S^d at the north pole, so we can't simply map it to the corresponding complex obtained from the polar dual of the lifted points $l(p_i)$ on the paraboloid $\mathcal{P}$. However, using projective completions, we can indeed define this mapping and recover the Voronoi diagram of P.

Definition 13.15. Given any set of points $P = \{p_1, \ldots, p_n\} \subseteq \mathbb{E}^d$, the *lifted Voronoi complex* associated with P is the polar dual (w.r.t. $S^d \subseteq \mathbb{R}^{d+1}$) $\mathcal{V}(P) = (\mathcal{DC}(P))^* \subseteq \mathbb{R}^{d+1}$ of the lifted Delaunay complex $\mathcal{DC}(P)$, and $\widetilde{\mathcal{V}}(P) \subseteq \mathbb{P}^{d+1}$ is the projective completion of $\mathcal{V}(P)$. See Figure 13.46. The polyhedral complex $\mathcal{V}or(P) = \varphi_{d+1}(\widetilde{\pi}_N(\widetilde{\mathcal{V}}(P)) \cap 2^{U_{d+1}}) \subseteq 2^{\mathbb{E}^d}$ is the *Voronoi complex of* P, or *Voronoi diagram of* P. See Figure 13.47.

As in Section 13.7 (just after Definition 13.12), it is easy to see that

$$\mathcal{V}or(P) = \varphi_{d+1}(\widetilde{\pi}_N(\widetilde{\mathcal{V}}(P)) \cap 2^{U_{d+1}}) = \pi_N(\mathcal{V}(P)).$$

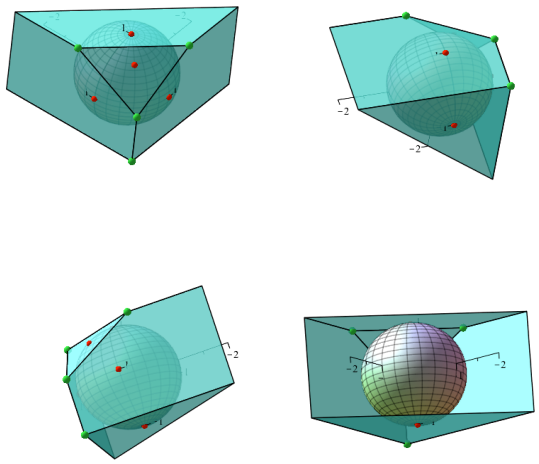


Figure 13.40: The polar dual $\mathcal{D}(P)^*$, where $P = \{(0,0,0),(1,0,0),(0,1,0),\ (2,2,0)\}$. The inverse stereographic projection of each point of P, along with N, is the depicted by a red dot, and the five teal faces of the unbounded wedge are tangent to each red point.

Definition 13.16. Given any set of points $P = \{p_1, \ldots, p_n\} \subseteq \mathbb{E}^d$, let $\mathcal{V}'(P) = (\mathcal{DC}'(P))^*$ be the polar dual (w.r.t. $\mathcal{P} \subseteq \mathbb{R}^{d+1}$) of the "standard" Delaunay complex of Definition 13.14, and let $\widetilde{\mathcal{V}}'(P) = \widetilde{\mathcal{DC}}'(P) \subseteq \mathbb{P}^d$ be its projective completion. The *standard Voronoi diagram* is given by $\mathcal{V}or'(P) = p_{d+1}(\mathcal{V}'(P))$; see Definition 17.2.7 of Boissonnat and Yvinec [12].

It is not hard to check that

$$\mathcal{V}or'(P) = p_{d+1}(\mathcal{V}'(P)) = \varphi_{d+1}(\widetilde{p}_{d+1}(\widetilde{\mathcal{V}}'(P)) \cap U_{d+1}).$$

In order to prove our second main theorem we need to show that θ has a good behavior with respect to tangent spaces. Recall from Section 12.2 that for any point $a = (a_1 : \cdots : a_{d+2}) \in \mathbb{P}^{d+1}$, the tangent hyperplane $T_a\widetilde{S^d}$ to the sphere $\widetilde{S^d}$ at a is given by the equation

$$\sum_{i=1}^{d+1} a_i x_i - a_{d+2}x_{d+2} = 0.$$

Similarly, the tangent hyperplane $T_a\widetilde{\mathcal{P}}$ to the paraboloid $\widetilde{\mathcal{P}}$ at a is given by the equation

$$2\sum_{i=1}^{d} a_i x_i - a_{d+2}x_{d+1} - a_{d+1}x_{d+2} = 0.$$

If we lift a point $a \in \mathbb{E}^d$ to $\widetilde{S^d}$ by $\widetilde{\tau}_N \circ \psi_{d+1}$ and to $\widetilde{\mathcal{P}}$ by $\widetilde{l}$, it turns out that the image of the tangent hyperplane to $\widetilde{S^d}$ at $\widetilde{\tau}_N \circ \psi_{d+1}(a)$ by θ is the tangent hyperplane to $\widetilde{\mathcal{P}}$ at $\widetilde{l}(a)$.

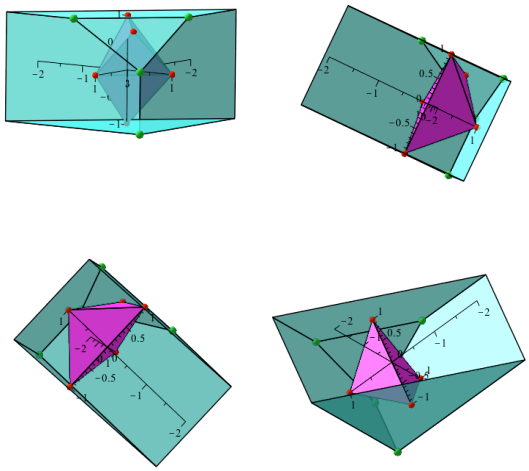


Figure 13.41: The polar dual $\mathcal{D}(P)^*$, where $P = \{(0,0,0),(1,0,0),(0,1,0),\ (2,2,0)\}$, and $\mathcal{D}(P)$ is the Delaunay polytope of Figures 13.29 and 13.30.

Proposition 13.12. *The map θ has the following properties:*

(1) For any point $a = (a_1, \ldots, a_d) \in \mathbb{E}^d$, we have

$$\theta(T_{\widetilde{\tau}_N \circ \psi_{d+1}(a)} \widetilde{S^d}) = T_{\widetilde{l}(a)} \widetilde{\mathcal{P}},$$

that is, θ preserves tangent hyperplanes.

(2) For every $(d-1)$-sphere $S \subseteq \mathbb{E}^d$, we have

$$\theta(\widetilde{\tau}_N \circ \psi_{d+1}(S)) = \widetilde{l}(\widetilde{S}),$$

that is, θ preserves lifted $(d-1)$-spheres.

Proof. (1) By Proposition 13.6, we know that

$$\widetilde{l} = \theta \circ \widetilde{\tau}_N \circ \psi_{d+1}$$

and we proved in Section 12.3 (Proposition 12.6) that projectivities preserve tangent spaces. Thus,

$$\theta(T_{\widetilde{\tau}_N \circ \psi_{d+1}(a)} \widetilde{S^d}) = T_{\theta \circ \widetilde{\tau}_N \circ \psi_{d+1}(a)} \theta(\widetilde{S^d}) = T_{\widetilde{l}(a)} \widetilde{\mathcal{P}},$$

as claimed.

(2) This follows immediately from the equation $\widetilde{l} = \theta \circ \widetilde{\tau}_N \circ \psi_{d+1}$. □

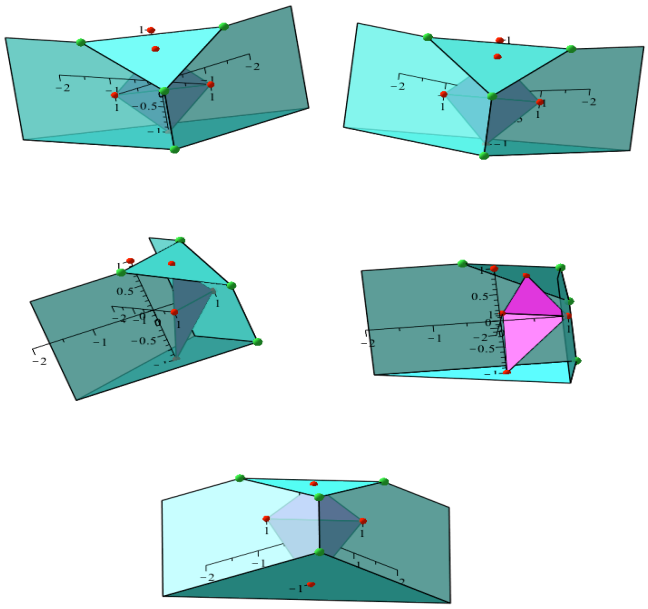


Figure 13.42: Five views of the dual to the lifted Delaunay complex of Figure 13.31.

Given any two distinct points $a = (a_1, \ldots, a_d)$ and $b = (b_1, \ldots, b_d)$ in $\mathbb{E}^d$, recall that the bisector hyperplane $H_{a,b}$ of a and b is given by

$$(b_1 - a_1)x_1 + \cdots + (b_d - a_d)x_d = (b_1^2 + \cdots + b_d^2)/2 - (a_1^2 + \cdots + a_d^2)/2.$$

We have the following useful proposition:

Proposition 13.13. *Given any two distinct points $a = (a_1, \ldots, a_d)$ and $b = (b_1, \ldots, b_d)$ in $\mathbb{E}^d$, the image under the projection $\widetilde{\pi}_N$ of the intersection $T_{\widetilde{\tau}_N \circ \psi_{d+1}(a)}\widetilde{S^d} \cap T_{\widetilde{\tau}_N \circ \psi_{d+1}(b)}\widetilde{S^d}$ of the tangent hyperplanes at the lifted points $\widetilde{\tau}_N \circ \psi_{d+1}(a)$ and $\widetilde{\tau}_N \circ \psi_{d+1}(b)$ on the sphere $\widetilde{S^d} \subseteq \mathbb{P}^{d+1}$ is the embedding of the bisector hyperplane $H_{a,b}$ of a and b into $\mathbb{P}^d$; that is,*

$$\widetilde{\pi}_N(T_{\widetilde{\tau}_N \circ \psi_{d+1}(a)}\widetilde{S^d} \cap T_{\widetilde{\tau}_N \circ \psi_{d+1}(b)}\widetilde{S^d}) = \psi_{d+1}(H_{a,b}).$$

Proof. In view of the geometric interpretation of $\widetilde{\pi}_N$ given earlier, we need to find the equation of the hyperplane H passing through the intersection of the tangent hyperplanes $T_{\widetilde{\tau}_N \circ \psi_{d+1}(a)}\widetilde{S^d}$ and $T_{\widetilde{\tau}_N \circ \psi_{d+1}(b)}\widetilde{S^d}$, and passing through the north pole, and then it is geometrically obvious that

$$\widetilde{\pi}_N(T_{\widetilde{\tau}_N \circ \psi_{d+1}(a)}\widetilde{S^d} \cap T_{\widetilde{\tau}_N \circ \psi_{d+1}(b)}\widetilde{S^d}) = H \cap H_{d+1}(0),$$

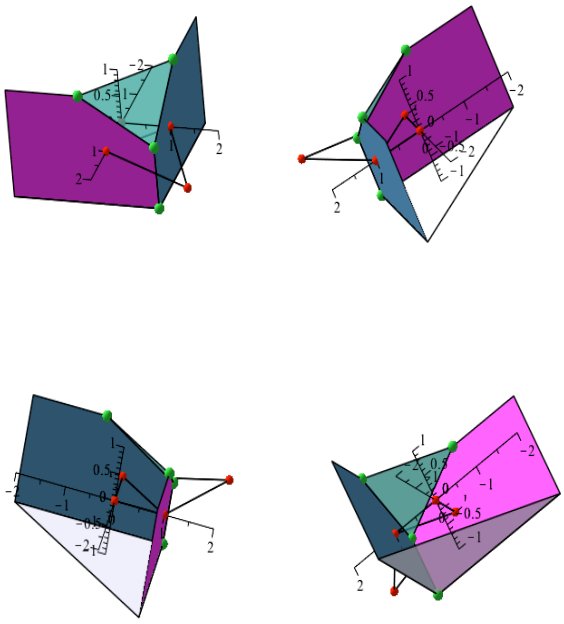


Figure 13.43: Four views of the relationship between $\mathcal{DC}(P)^*$ of Figure 13.42 and the Delaunay complex of Figure 13.32.

where $H_{d+1}(0)$ is the hyperplane (in $\mathbb{P}^{d+1}$) of equation $x_{d+1} = 0$. Recall that $T_{\widetilde{\tau}_N \circ \psi_{d+1}(a)}\widetilde{S^d}$ and $T_{\widetilde{\tau}_N \circ \psi_{d+1}(b)}\widetilde{S^d}$ are given by

$$E_1 = 2\sum_{i=1}^{d} a_i x_i + (\sum_{i=1}^{d} a_i^2 - 1)x_{d+1} - (\sum_{i=1}^{d} a_i^2 + 1)x_{d+2} = 0$$

and

$$E_2 = 2\sum_{i=1}^{d} b_i x_i + (\sum_{i=1}^{d} b_i^2 - 1)x_{d+1} - (\sum_{i=1}^{d} b_i^2 + 1)x_{d+2} = 0.$$

The hyperplanes passing through $T_{\widetilde{\tau}_N \circ \psi_{d+1}(a)}\widetilde{S^d} \cap T_{\widetilde{\tau}_N \circ \psi_{d+1}(b)}\widetilde{S^d}$ are given by an equation of the form

$$\lambda E_1 + \mu E_2 = 0,$$

with $\lambda, \mu \in \mathbb{R}$. Furthermore, in order to contain the north pole, this equation must vanish for $x = (0\colon \cdots \colon 0\colon 1\colon 1)$. But, observe that setting $\lambda = -1$ and $\mu = 1$ gives a solution since the corresponding equation is

$$2\sum_{i=1}^{d}(b_i - a_i)x_i + (\sum_{i=1}^{d} b_i^2 - \sum_{i=1}^{d} a_i^2)x_{d+1} - (\sum_{i=1}^{d} b_i^2 - \sum_{i=1}^{d} a_i^2)x_{d+2} = 0,$$

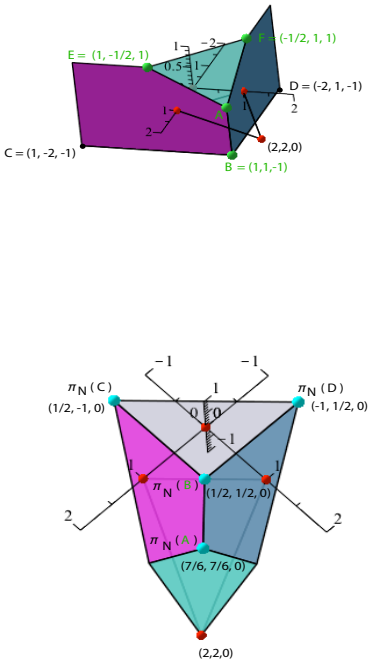


Figure 13.44: The Voronoi diagram for the Delaunay triangulation of the red dots $P = \{(0,0,0),(1,0,0),(0,1,0),(2,2,0)\}$. The green and black dots are projected, via π_N, onto the aqua dots. Note that E and F are mapped to infinity.

and it vanishes on $(0\colon \cdots \colon 0\colon 1\colon 1)$. But then, the intersection of H with the hyperplane $H_{d+1}(0)$ of equation $x_{d+1} = 0$ is given by

$$2\sum_{i=1}^{d}(b_i - a_i)x_i - (\sum_{i=1}^{d} b_i^2 - \sum_{i=1}^{d} a_i^2)x_{d+2} = 0.$$

Since we view $\mathbb{P}^d$ as the hyperplane $H_{d+1}(0) \subseteq \mathbb{P}^{d+1}$ and since the coordinates of points in $H_{d+1}(0)$ are of the form $(x_1\colon \cdots \colon x_d\colon 0\colon x_{d+2})$, the above equation is equivalent to the equation of $\psi_{d+1}(H_{a,b})$ in $\mathbb{P}^d$ in which x_{d+1} is replaced by x_{d+2}. □

Here is the second main theorem of this chapter.

Theorem 13.14. *Given any set of points $P = \{p_1, \ldots, p_n\} \subseteq \mathbb{E}^d$, we have*

$$\theta(\widetilde{\mathcal{V}}(P)) = \widetilde{\mathcal{V}}'(P)$$

and

$$\mathcal{V}or(P) = \mathcal{V}or'(P).$$

Therefore, the two notions of Voronoi diagrams agree.

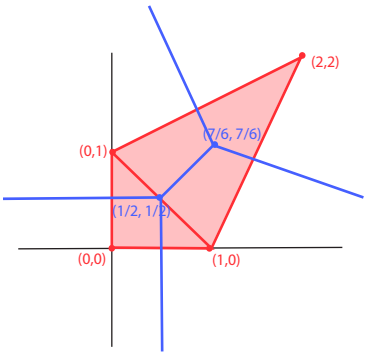


Figure 13.45: Another view of the Voronoi diagram (in blue) for the Delaunay triangulation (in red) of $P = \{(0,0,0),(1,0,0),(0,1,0),(2,2,0)\}$.

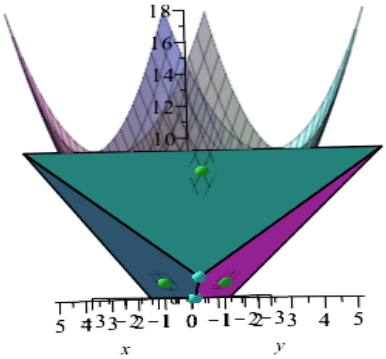


Figure 13.46: The polar dual of $\mathcal{DC}'(P)$ from Figure 13.38.

Proof. By definition,

$$\widetilde{\mathcal{V}}(P) = \widetilde{\mathcal{V}(P)} = \widetilde{\mathcal{DC}(P)^*},$$

and by Proposition 12.13,

$$\widetilde{\mathcal{DC}(P)^*} = \left(\widetilde{\mathcal{DC}(P)}\right)^* = (\widetilde{\mathcal{DC}}(P))^*,$$

so

$$\widetilde{\mathcal{V}}(P) = (\widetilde{\mathcal{DC}}(P))^*.$$

By Proposition 12.11,

$$\theta(\widetilde{\mathcal{V}}(P)) = \theta((\widetilde{\mathcal{DC}}(P))^*) = (\theta(\widetilde{\mathcal{DC}}(P)))^*,$$

and by Theorem 13.10,

$$\theta(\widetilde{\mathcal{DC}}(P)) = \widetilde{\mathcal{DC}}'(P),$$

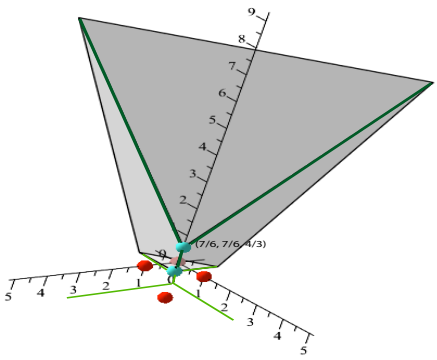


Figure 13.47: The orthogonal projection of the dark green lines onto the lighter green lines (which lie in the xy-plane, provide the Voronoi diagram of $P = \{(0,0,0),(1,0,0),(2,2,0),(0,1,0)\}$. This is precisely the same Voronoi diagram as in Figure 13.45.

so we get

$$\theta(\widetilde{\mathcal{V}}(P)) = (\widetilde{\mathcal{DC}}'(P))^*.$$

But, by Proposition 12.13 again,

$$(\widetilde{\mathcal{DC}}'(P))^* = \left(\widetilde{\mathcal{DC}'(P)}\right)^* = \widetilde{\mathcal{DC}'(P)^*} = \widetilde{\mathcal{V}'(P)} = \widetilde{\mathcal{V}}'(P).$$

Therefore,

$$\theta(\widetilde{\mathcal{V}}(P)) = \widetilde{\mathcal{V}}'(P),$$

as claimed.

As $\widetilde{\pi}_N = \widetilde{p}_{d+1} \circ \theta$ by Proposition 13.6, we get

$$\begin{aligned}
\mathcal{V}or'(P) &= \varphi_{d+1}(\widetilde{p}_{d+1}(\widetilde{\mathcal{V}}'(P)) \cap 2^{U_{d+1}}) \\
&= \varphi_{d+1}(\widetilde{p}_{d+1} \circ \theta(\widetilde{\mathcal{V}}(P)) \cap 2^{U_{d+1}}) \\
&= \varphi_{d+1}(\widetilde{\pi}_N(\widetilde{\mathcal{V}}(P)) \cap 2^{U_{d+1}}) \\
&= \mathcal{V}or(P),
\end{aligned}$$

finishing the proof. □

We can also prove the proposition below which shows directly that $\mathcal{V}or(P)$ is the Voronoi diagram of P. Recall that that $\widetilde{\mathcal{V}}(P)$ is the projective completion of $\mathcal{V}(P)$. We observed in Section 12.2 (see page 312) that in the patch U_{d+1}, there is a bijection between the faces of $\widetilde{\mathcal{V}}(P)$ and the faces of $\mathcal{V}(P)$. Furthermore, the projective completion $\widetilde{H}$ of every hyperplane $H \subseteq \mathbb{R}^d$ is also a hyperplane, and it is easy to see that if H is tangent to $\mathcal{V}(P)$, then $\widetilde{H}$ is tangent to $\widetilde{\mathcal{V}}(P)$.

Proposition 13.15. *Given any set of points $P = \{p_1, \ldots, p_n\} \subseteq \mathbb{E}^d$, for every $p \in P$, if F is the facet of $\mathcal{V}(P)$ that contains $\tau_N(p)$, if H is the tangent hyperplane at $\tau_N(p)$ to S^d, and if F is cut out by the hyperplanes $H, H_1, \ldots, H_{k_p}$, in the sense that*

$$F = (H \cap H_1)_- \cap \cdots \cap (H \cap H_{k_p})_-,$$

where $(H \cap H_i)_-$ denotes the closed half-space in H containing $\tau_N(p)$ determined by $H \cap H_i$, then

$$V(p) = \varphi_{d+1}(\widetilde{\pi}_N(\widetilde{H} \cap \widetilde{H}_1)_- \cap \cdots \cap \widetilde{\pi}_N(\widetilde{H} \cap \widetilde{H}_{k_p})_- \cap U_{d+1})$$

is the Voronoi region of p (where $\varphi_{d+1}(\widetilde{\pi}_N(\widetilde{H} \cap \widetilde{H}_i)_- \cap U_{d+1})$ is the closed half-space containing p). If P is in general position and $\dim(\mathrm{conv}(P)) = d$, then $\mathcal{V}(P)$ and $\mathcal{V}or(P)$ are simple polyhedra (every vertex belongs to $d+1$ facets).

Proof. Recall that by Proposition 13.6,

$$\widetilde{\tau}_N \circ \psi_{d+1} = \psi_{d+2} \circ \tau_N.$$

Each $H_i = T_{\tau_N(p_i)} S^d$ is the tangent hyperplane to S^d at $\tau_N(p_i)$, for some $p_i \in P$. Now, by definition of the projective completion, the embedding $\mathcal{V}(P) \longrightarrow \widetilde{\mathcal{V}}(P)$ is given by $a \mapsto \psi_{d+2}(a)$. Thus, every point $p \in P$ is mapped to the point $\psi_{d+2}(\tau_N(p)) = \widetilde{\tau_N}(\psi_{d+1}(p))$, and we also have $\widetilde{H}_i = T_{\widetilde{\tau}_N \circ \psi_{d+1}(p_i)} S^d$ and $\widetilde{H} = T_{\widetilde{\tau}_N \circ \psi_{d+1}(p)} S^d$. By Proposition 13.13,

$$\widetilde{\pi}_N(T_{\widetilde{\tau}_N \circ \psi_{d+1}(p)} S^d \cap T_{\widetilde{\tau}_N \circ \psi_{d+1}(p_i)} S^d) = \psi_{d+1}(H_{p,p_i})$$

is the embedding of the bisector hyperplane of p and p_i in $\mathbb{P}^d$, so the first part holds.

Since $\dim(\mathrm{conv}(P)) = d$ every vertex of $\mathcal{V}(P)$ must belong to at least $d+1$ faces.Now, assume that some vertex $v \in \mathcal{V}(P) = \mathcal{DC}(P)^*$ belongs to $k \geq d+2$ facets of $\mathcal{V}(P)$. By polar duality, this means that the facet F dual of v has $k \geq d+2$ vertices $\tau_N(p_1), \ldots, \tau_N(p_k)$ of $\mathcal{DC}(P)$. However, this contradicts Proposition 13.11. The fact that $\mathcal{V}or(P)$ is a simple polyhedron was aready proved in Proposition 13.1. □

Note that if $m = \dim(\mathrm{conv}(P)) < d$, then the Voronoi complex $\mathcal{V}(P)$ may not have any vertices.

We conclude our presentation of Voronoi diagrams and Delaunay triangulations with a short section on applications.

13.9 Applications of Voronoi Diagrams and Delaunay Triangulations

The examples below are taken from O'Rourke [46]. Other examples can be found in Preparata and Shamos [49], Boissonnat and Yvinec [12], and de Berg, Van Kreveld, Overmars, and Schwarzkopf [6].

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 [46]). 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 [46]).

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 [46].

Acknowledgement. I wish to thank Marcelo Siqueira for suggesting many improvements and for catching many bugs with his "eagle eye."

Bibliography

[1] P.S. Alexandrov. *Combinatorial Topology*. Dover, first edition, 1998. Three volumes bound as one.

[2] Noga Alon and Gil Kalai. A simple proof of the upper-bound theorem. *European J. Comb.*, 6:211–214, 1985.

[3] Michael Artin. *Algebra*. Prentice Hall, first edition, 1991.

[4] Alexander Barvinok. *A Course in Convexity*. GSM, Vol. 54. AMS, first edition, 2002.

[5] Margaret M. Bayer and Carl W. Lee. Combinatorial aspects of convex polytopes. In P.M. Gruber and J.M. Wills, editors, *Handbook of Convex Geometry*, pages 485–534. Elsevier Science, 1993.

[6] M. Berg, M. Van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry. Algorithms and Applications*. Springer, first edition, 1997.

[7] Marcel Berger. *Géométrie 1*. Nathan, 1990. English edition: Geometry 1, Universitext, Springer Verlag.

[8] Marcel Berger. *Géométrie 2*. Nathan, 1990. English edition: Geometry 2, Universitext, Springer Verlag.

[9] Dimitri P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, first edition, 2009.

[10] Dimitris Bertsimas and John N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, third edition, 1997.

[11] J. Billera, Louis and Anders Björner. Faces numbers of polytopes and complexes. In J.E. Goodman and Joe O'Rourke, editors, *Handbook of Discrete and Computational Geometry*, pages 291–310. CRC Press, 1997.

[12] J.-D. Boissonnat and M. Yvinec. *Géométrie Algorithmique*. Ediscience International, first edition, 1995.

[13] Nicolas Bourbaki. *Espaces Vectoriels Topologiques*. Eléments de Mathématiques. Masson, 1981.

[14] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, first edition, 2004.

[15] David A. Brannan, Matthew F. Esplen, and Jeremy J. Gray. *Geometry*. Cambridge University Press, first edition, 1999.

[16] K.Q. Brown. Voronoi diagrams from convex hulls. *Inform. Process. Lett.*, 9:223–228, 1979.

[17] Heinz Bruggesser and Peter Mani. Shellable decompositions of cells and spheres. *Math. Scand.*, 29:197–205, 1971.

[18] Vasek Chvatal. *Linear Programming*. W.H. Freeman, first edition, 1983.

[19] P.G. Ciarlet. *Introduction to Numerical Matrix Analysis and Optimization*. Cambridge University Press, first edition, 1989. French edition: Masson, 1994.

[20] H.S.M. Coxeter. *Regular Polytopes*. Dover, third edition, 1973.

[21] H.S.M. Coxeter. *Introduction to Geometry*. Wiley, second edition, 1989.

[22] Peter Cromwell. *Polyhedra*. Cambridge University Press, first edition, 1994.

[23] G.L. Dirichlet. Über die reduktion der positiven quadratischen formen mid drei unbestimmten ganzen zahlen. *Journal für die reine und angewandte Mathematik*, 40:209–227, 1850.

[24] H. Edelsbrunner and R. Seidel. Voronoi diagrams and arrangements. *Discrete Computational Geometry*, 1:25–44, 1986.

[25] Herbert Edelsbrunner. *Geometry and Topology for Mesh Generation*. Cambridge University Press, first edition, 2001.

[26] Günter Ewald. *Combinatorial Convexity and Algebraic Geometry*. GTM No. 168. Springer Verlag, first edition, 1996.

[27] Jean Fresnel. *Méthodes Modernes En Géométrie*. Hermann, first edition, 1998.

[28] William Fulton. *Introduction to Toric Varieties*. Annals of Mathematical Studies, No. 131. Princeton University Press, 1997.

[29] Jean H. Gallier. *Curves and Surfaces In Geometric Modeling: Theory And Algorithms*. Morgan Kaufmann, 1999.

[30] Jean H. Gallier. *Geometric Methods and Applications, For Computer Science and Engineering*. TAM, Vol. 38. Springer, second edition, 2011.

[31] Jean H. Gallier. Notes on Convex Sets, Polytopes, Polyhedra, Combinatorial Topology, Voronoi Diagrams, and Delaunay Triangulations. Technical report, University of Pennsylvania, CIS Department, Philadelphia, PA 19104, 2016. Book in Preparation.

[32] E.N. Gilbert. Random subdivisions of space into crystals. *Annals of Math. Stat.*, 33:958–972, 1962.

[33] Jacob E. Goodman and Joseph O'Rourke. *Handbook of Discrete and Computational Geometry*. CRC Press, second edition, 2004.

[34] R. Graham and F. Yao. A whirlwind tour of computational geometry. *American Mathematical Monthly*, 97(8):687–701, 1990.

[35] Donald T. Greenwood. *Principles of Dynamics*. Prentice Hall, second edition, 1988.

[36] Branko Grünbaum. *Convex Polytopes*. GTM No. 221. Springer Verlag, second edition, 2003.

[37] D. Hilbert and S. Cohn-Vossen. *Geometry and the Imagination*. Chelsea Publishing Co., 1952.

[38] Serge Lang. *Algebra*. Addison Wesley, third edition, 1993.

[39] Serge Lang. *Real and Functional Analysis*. GTM 142. Springer Verlag, third edition, 1996.

[40] Peter Lax. *Functional Analysis*. Wiley, first edition, 2002.

[41] Jiri Matousek. *Lectures on Discrete Geometry*. GTM No. 212. Springer Verlag, first edition, 2002.

[42] Jiri Matousek and Bernd Gartner. *Understanding and Using Linear Programming*. Universitext. Springer Verlag, first edition, 2007.

[43] Peter McMullen. The maximum number of faces of a convex polytope. *Mathematika*, 17:179–184, 1970.

[44] T. Molla. Class notes, math 588 example 5. Technical report, 2015. http://myweb.usf.edu/molla/2015_spring_math588/example5.pdf.

[45] James R. Munkres. *Elements of Algebraic Topology*. Addison-Wesley, first edition, 1984.

[46] Joseph O'Rourke. *Computational Geometry in C*. Cambridge University Press, second edition, 1998.

[47] Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization. Algorithms and Complexity*. Dover, first edition, 1998.

[48] Dan Pedoe. *Geometry, A comprehensive Course*. Dover, first edition, 1988.

[49] F.P. Preparata and M.I. Shamos. *Computational Geometry: An Introduction*. Springer Verlag, first edition, 1988.

[50] J.-J. Risler. *Mathematical Methods for CAD*. Masson, first edition, 1992.

[51] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, 1970.

[52] Pierre Samuel. *Projective Geometry*. Undergraduate Texts in Mathematics. Springer Verlag, first edition, 1988.

[53] Alexander Schrijver. *Theory of Linear and Integer Programming*. Wiley, first edition, 1999.

[54] Raimund Seidel. The upper-bound theorem for polytopes: an easy proof of its asymptotic version. *Comput. Geometry: Theory and Applications*, 5:115–116, 1995.

[55] Ernst Snapper and Robert J. Troyer. *Metric Affine Geometry*. Dover, first edition, 1989.

[56] John Stallings. *Lectures on Polyhedral Topology*. Tata Institute, first edition, 1967.

[57] Richard P. Stanley. The number of faces of simplicial polytopes and spheres. In J.E Goodman, E. Lutwak, J. Malkevitch, and P. Pollack, editors, *Discrete Geometry and Convexity*, pages 212–223. Annals New York Academy of Sciences, 1985.

[58] Richard P. Stanley. *Combinatorics and Commutative Algebra*. Progress in Mathematics, No. 41. Birkhäuser, second edition, 1996.

[59] J. Stolfi. *Oriented Projective Geometry*. Academic Press, first edition, 1991.

[60] Gilbert Strang. *Linear Algebra and its Applications*. Saunders HBJ, third edition, 1988.

[61] Bernd Sturmfels. *Gröbner Bases and Convex Polytopes*. ULS, Vol. 8. AMS, first edition, 1996.

[62] Rekha R. Thomas. *Lectures in Geometric Combinatorics*. STML, Vol. 33. AMS, first edition, 2006.

[63] P. Thurston, William. *Three-Dimensional Geometry and Topology, Vol. 1*. Princeton University Press, first edition, 1997. Edited by Silvio Levy.

[64] Claude Tisseron. *Géométries affines, projectives, et euclidiennes*. Hermann, first edition, 1994.

[65] Frederick A. Valentine. *Convex Sets*. McGraw–Hill, first edition, 1964.

[66] Robert J. Vanderbei. *Linear Programming: Foundations and Extensions*. Springer, fourth edition, 2014.

[67] Lucas Vienne. *Présentation algébrique de la géométrie classique*. Vuibert, first edition, 1996.

[68] M.G. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *J. Reine u. Agnew. Math.*, 134:198–287, 1908.

[69] Gunter Ziegler. *Lectures on Polytopes*. GTM No. 152. Springer Verlag, first edition, 1997.