

Mathematical Visualization

Jan Techter

September 22, 2025

Lecture notes for Mathematical Visualization 1, Summer 2025.

Contents

1	Projective geometry	3
1.1	Some motivation: Incidences between points and lines	3
1.2	Definition of projective spaces	5
1.3	Homogeneous coordinates on \mathbb{RP}^n	5
1.4	Projective subspaces	6
1.5	Meet and join	7
1.6	Desargues' theorem	7
1.7	Duality	9
1.8	Projective transformations	11
2	Plane curves and envelopes of lines	16
2.1	Plane curves	16
2.2	Discrete plane curves	18
2.3	Envelopes	20
2.4	Evolute	22
2.5	Involute	24
3	Conics and quadrics	27
3.1	Bilinear forms	28
3.2	Quadrics	28
3.3	Projective classification of quadrics in \mathbb{RP}^n	30
3.4	Affine classification of quadrics in $\mathbb{R}^n \subset \mathbb{RP}^n$	35
3.5	Signature of subspaces	36
3.6	Tangent lines and tangent cones	37
3.7	Polarity and tangent planes	38
4	Pencils of quadrics	40
4.1	Classification of pencils of conics	43
4.2	Classification of pencils of quadrics	46

5	Fractals	51
5.1	Iterated function systems	52
5.2	Fractal dimensions	55
5.3	Iteration of complex functions	68
6	Möbius geometry	80
6.1	The elementary model of Möbius geometry	80
6.2	Two-dimensional Möbius geometry	82
6.3	Schottky groups and limit sets	87

1 Projective geometry

1.1 Some motivation: Incidences between points and lines

The elementary figures of projective geometry are points, straight lines, and planes. The elementary results of projective geometry deal with the simplest possible relations between these entities, namely their *incidence*. The word incidence covers all the following relations: A point lying on a straight line, a point lying in a plane, a straight line lying in a plane. Clearly, the three statements that a straight line passes through a point, that a plane passes through a point, that a plane passes through a straight line, are respectively equivalent to the first three. The term incidence was introduced to give these three pairs of statements symmetrical form: a straight line is incident with a point, a plane is incident with a point, a plane is incident with a straight line. (Geometry and the Imagination – Hilbert, Cohn-Vossen)

In projective geometry, we are interested in statements and configurations that are invariant under *projective transformations*. E.g., the incidence of a point lying on a line is invariant under projection from one plane to another (from some point). Let us take a closer look at this incidence in the plane.

A point in the Euclidean plane \mathbb{R}^2 can be described by two Cartesian coordinates

$$p = (p_1, p_2) \in \mathbb{R}^2,$$

and a line by

$$\ell = \{p = (p_1, p_2) \in \mathbb{R}^2 \mid \langle n, p \rangle + h = 0\}$$

with some $n = (n_1, n_2) \in \mathbb{S}^1$ and $h \in \mathbb{R}$, where n can be interpreted as the unit normal vector of ℓ and h as the oriented distance of the origin to ℓ .

Note that the equation for ℓ can be multiplied by any scalar $\lambda \in \mathbb{R}$, $\lambda \neq 0$ without changing the line. Thus, we can replace (n_1, n_2, h) by

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \lambda \begin{pmatrix} n_1 \\ n_2 \\ h \end{pmatrix}, \quad \text{with some } \lambda \in \mathbb{R}, \lambda \neq 0,$$

and write the equation for the line as

$$a_1 p_1 + a_2 p_2 + a_3 = \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ 1 \end{pmatrix} = 0$$

Similarly, we can replace $(p_1, p_2, 1)$ by any non-zero scalar multiple

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \mu \begin{pmatrix} p_1 \\ p_2 \\ 1 \end{pmatrix}, \quad \text{with some } \mu \in \mathbb{R}, \mu \neq 0,$$

from which the Cartesian coordinates of p can be recovered by

$$p_1 = \frac{x_1}{x_3}, \quad p_2 = \frac{x_2}{x_3}.$$

The triple (x_1, x_2, x_3) , and in particular $(p_1, p_2, 1)$, are called *homogeneous coordinates* of p .

Now the equation of the *incidence* of the point p lying on the line ℓ ($p \in \ell$), or equivalently, the line ℓ passing through the point p ($\ell \ni p$) has the symmetric form

$$a_1x_1 + a_2x_2 + a_3x_3 = \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = 0 \quad (1)$$

Example 1.1. How to determine if three points $p, q, r \in \mathbb{R}^2$ lie on a line?

Equation (1) is a linear homogeneous equation in (a_1, a_2, a_3) . Thus, there exists a line passing through these three points if and only if the linear homogeneous system

$$\begin{pmatrix} p_1 & p_2 & 1 \\ q_1 & q_2 & 1 \\ r_1 & r_2 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = 0$$

has a non-trivial solution, which is equivalent to

$$\det \begin{pmatrix} p_1 & p_2 & 1 \\ q_1 & q_2 & 1 \\ r_1 & r_2 & 1 \end{pmatrix} = 0.$$

Example 1.2. How to compute the intersection point of two lines?

$$\begin{aligned} \ell &= \{p \in \mathbb{R}^2 \mid a_1p_1 + a_2p_2 + a_3 = 0\} \\ \tilde{\ell} &= \{p \in \mathbb{R}^2 \mid \tilde{a}_1p_1 + \tilde{a}_2p_2 + \tilde{a}_3 = 0\} \end{aligned}$$

Its homogeneous coordinates are given by a solution of the linear homogeneous system

$$\begin{pmatrix} a_1 & a_2 & a_3 \\ \tilde{a}_1 & \tilde{a}_2 & \tilde{a}_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0. \quad (2)$$

If we assume that the two lines are distinct, i.e., the two rows are independent, then the solution space is one-dimensional

$$\text{span}\{x\} = \{\lambda x \mid \lambda \in \mathbb{R}\} \quad \text{with some } x \in \mathbb{R}^3, x \neq 0,$$

and we obtain the intersection point $p \in \mathbb{R}^2$ with

$$p_1 = \frac{x_1}{x_3}, \quad p_2 = \frac{x_2}{x_3}.$$

What if $x_3 = 0$? Then

$$\det \begin{pmatrix} a_1 & a_2 \\ \tilde{a}_1 & \tilde{a}_2 \end{pmatrix} = 0$$

and thus ℓ and $\tilde{\ell}$ are parallel.

The linear homogeneous system (2) always has a solution. Thus, in homogeneous coordinates of the plane two lines always intersect. In particular, for two parallel lines, the point of intersection has homogeneous coordinates of the form $(x_1, x_2, 0)$ which represents a point not in \mathbb{R}^2 , but “at infinity”.

1.2 Definition of projective spaces

Let V be a vector space of dimension $n + 1$ over a field \mathbb{F} . Then the *projective space* of V is the set

$$P(V) := \{1\text{-dimensional subspaces of } V\}$$

Its dimension is given by

$$\dim P(V) := \dim V - 1 = n.$$

For $x \in V \setminus \{0\}$ we write $[x] := \text{span}\{x\}$. Then $[x]$ is a point in $P(V)$, and x is called a *representative vector* for this point.

If $\lambda \in \mathbb{F} \setminus \{0\}$ then $[\lambda x] = [x]$, and λx is another representative vector for the same point. This defines an equivalence relation on $V \setminus \{0\}$

$$x \sim y \iff x = \lambda y, \quad \text{for some } \lambda \in \mathbb{F} \setminus \{0\},$$

and we can identify

$$P(V) \cong (V \setminus \{0\}) / \sim.$$

For now we will only consider the real projective space

$$\mathbb{RP}^n := P(\mathbb{R}^{n+1}).$$

1.3 Homogeneous coordinates on \mathbb{RP}^n

For a point $[x_1, \dots, x_{n+1}] \in \mathbb{RP}^n$ the coordinates of a representative vector $(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1}$ are called *homogeneous coordinates*. They are unique up to a common scalar multiple

$$[x_1, \dots, x_{n+1}] = [\lambda x_1, \dots, \lambda x_{n+1}]$$

for $\lambda \in \mathbb{R} \setminus \{0\}$.

If $x_{n+1} \neq 0$ then

$$[x_1, \dots, x_{n+1}] = \left[\frac{x_1}{x_{n+1}}, \dots, \frac{x_n}{x_{n+1}}, 1 \right] = [y_1, \dots, y_n, 1],$$

and (y_1, \dots, y_n) are called *affine coordinates* of the point $[x]$. This yields a decomposition of \mathbb{RP}^n into an affine part and a *hyperplane at infinity*

$$\mathbb{RP}^n = \underbrace{\{[x_1, \dots, x_{n+1}] \mid x_{n+1} \neq 0\}}_{\simeq \mathbb{R}^n} \cup \underbrace{\{[x_1, \dots, x_{n+1}] \mid x_{n+1} = 0\}}_{\simeq \mathbb{RP}^{n-1}}.$$

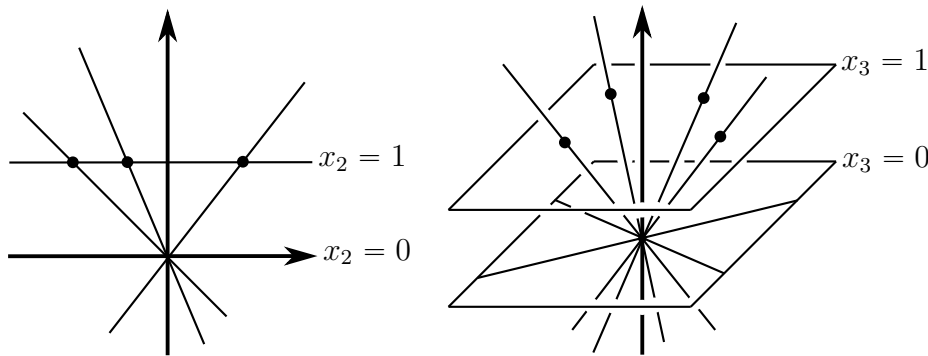


Figure 1. Affine coordinates for \mathbb{RP}^1 and \mathbb{RP}^2 .

Example 1.3 (The real projective line \mathbb{RP}^1). For the real projective line this decomposition is given by

$$\mathbb{RP}^1 \cong \mathbb{R} \cup \mathbb{RP}^0 = \mathbb{R} \cup \{\infty\},$$

where \mathbb{RP}^0 consists of only one point $[1, 0]$, which is usually denoted by ∞ , and allowed as an “admissible” affine coordinate.

Example 1.4 (The real projective plane \mathbb{RP}^2). For the real projective plane this decomposition is given by

$$\mathbb{RP}^2 \cong \mathbb{R}^2 \cup \mathbb{RP}^1.$$

Thus, we obtain the Euclidean plane compactified by a (projective) line at infinity.

Example 1.5 (The real projective 3-space \mathbb{RP}^3). For the real projective plane this decomposition is given by

$$\mathbb{RP}^3 \cong \mathbb{R}^3 \cup \mathbb{RP}^2.$$

Thus, we obtain the Euclidean 3-space compactified by a (projective) plane at infinity.

More generally, let b_1, \dots, b_{n+1} be a basis of \mathbb{R}^{n+1} . For $x \in \mathbb{R}^{n+1}$ let $x_1, \dots, x_{n+1} \in \mathbb{R}$ such that

$$x = \sum_{i=1}^{n+1} x_i b_i.$$

Then (x_1, \dots, x_{n+1}) are called *homogeneous coordinates* of the point $[x] \in \mathbb{RP}^n$ (with respect to b_1, \dots, b_{n+1}). They depend on the chosen basis and are unique up to a common scalar multiple. We then identify

$$[x] \cong [x_1, \dots, x_{n+1}].$$

A change of basis acts on the homogeneous coordinates as a general linear transformation

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{n+1} \end{bmatrix} \mapsto \begin{bmatrix} A \begin{pmatrix} x_1 \\ \vdots \\ x_{n+1} \end{pmatrix} \end{bmatrix}$$

with $A \in \text{GL}(\mathbb{R}^{n+1})$.

1.4 Projective subspaces

For a $(k + 1)$ -dimensional linear subspace $U \subset \mathbb{R}^{n+1}$ its projective space

$$\text{P}(U) \subset \mathbb{RP}^n$$

is called a k -dimensional *projective subspace* of \mathbb{RP}^n .

$\dim \text{P}(U)$	name
0	point
1	line
2	plane
k	k -plane
$n - 1$	hyperplane

Table 1. Naming conventions for projective (sub)spaces.

1.5 Meet and join

Let $P(U_1), P(U_2) \subset \mathbb{RP}^n$ be two projective subspaces. Then their intersection, or *meet*, is given by

$$P(U_1) \cap P(U_2) = P(U_1 \cap U_2),$$

and their span, or *join*, is given by

$$P(U_1) \vee P(U_2) = P(U_1 + U_2).$$

The dimension formula for linear subspaces carries over to projective subspaces:

$$\dim(P(U_1) \vee P(U_2)) + \dim(P(U_1) \cap P(U_2)) = \dim P(U_1) + \dim P(U_2).$$

In particular, a k_1 -plane and a k_2 -plane in an n -dimensional projective space with $k_1 + k_2 \geq n$ always intersect in an at least $(k_1 + k_2 - n)$ -dimensional projective subspace. Thus, certain incidences are always guaranteed in a projective space.

Example 1.6 (\mathbb{RP}^2). In \mathbb{RP}^2 two (distinct) lines always intersect in a point. In affine coordinates, the two lines are parallel if and only if the intersection point lies on the line at infinity.

Example 1.7 (\mathbb{RP}^3). In \mathbb{RP}^3 two (distinct) planes always intersect in a line. In affine coordinates, the two planes are parallel if and only if the intersection line lies in the plane at infinity.

However, in \mathbb{RP}^3 , two lines do not always intersect. They intersect if and only if they lie in a plane. In affine coordinates, two lines are parallel if and only if the intersection point lies in the plane at infinity.

1.6 Desargues' theorem

An *incidence theorem* is a statement about a projective configuration (of e.g. projective subspaces) where a certain set of incidences implies another set of incidences. As an example we state the theorem of Desargues. First in \mathbb{RP}^3 where it is very easy to verify, and then in \mathbb{RP}^2 .

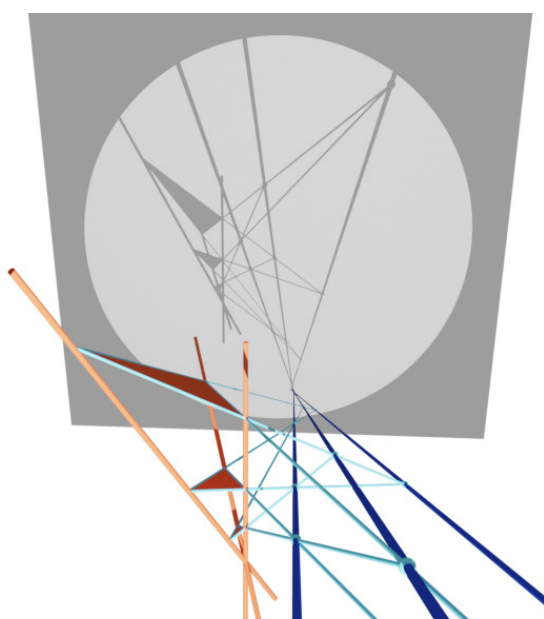


Figure 2. Three triangles in perspective and their shadow.

Theorem 1.1 (Desargues). *Let A, A', B, B', C, C' be six points in \mathbb{RP}^3 , such that A, B, C span a plane, and A', B', C' span another plane.*

If the three lines $AA', BB',$ and CC' are distinct and pass through a common point, then the three points $A'' = BC \cap B'C', B'' = CA \cap C'A',$ and $C'' = AB \cap A'B'$ lie on a common line.

Proof. First, the statement contains the implicit claim, that, e.g., the lines BC and $B'C'$ intersect in a point. Indeed, the four points B, C, B', C' lie in a plane since BB' and CC' are concurrent. Thus, the point $A'' = BC \cap B'C'$ exists.

The two planes

$$E = A \vee B \vee C, \quad E' = A' \vee B' \vee C'$$

intersect in a line $\ell = E \cap E'$. Since $BC \in E$ and $B'C' \in E'$, their intersection point A'' lies in ℓ . Similarly, $B'', C'' \in \ell$. \square

Consider what happens if we project such a configuration in \mathbb{RP}^3 from a point into a plane, and denote the image points by $\tilde{A}, \tilde{B}, \tilde{C}, \dots$. Then we obtain again six points $\tilde{A}, \tilde{A}', \tilde{B}, \tilde{B}', \tilde{C}, \tilde{C}'$ that satisfy that the lines $\tilde{A}\tilde{A}', \tilde{B}\tilde{B}',$ and $\tilde{C}\tilde{C}'$ are concurrent and that the points $\tilde{A}'' = \tilde{B}\tilde{C} \cap \tilde{B}'\tilde{C}', \tilde{B}'' = \tilde{C}\tilde{A} \cap \tilde{C}'\tilde{A}',$ and $\tilde{C}'' = \tilde{A}\tilde{B} \cap \tilde{A}'\tilde{B}'$ are collinear.

Indeed, Desargues theorem also holds in \mathbb{RP}^2 which can be shown by lifting it to \mathbb{RP}^3 .

Theorem 1.2. *Let A, A', B, B', C, C' be six points in \mathbb{RP}^2 , such that no three lie on a line.*

If the three lines $AA', BB',$ and CC' pass through a common point, then the three points $A'' = BC \cap B'C', B'' = CA \cap C'A',$ and $C'' = AB \cap A'B'$ lie on a common line.

Proof. We embed \mathbb{RP}^2 into \mathbb{RP}^3 as the plane $\mathbb{RP}^2 \cong E \subset \mathbb{RP}^3$. Thus, E is the plane which contains the two triangles $ABC, A'B'C'$, and the point P which is incident with the three lines $AA', BB',$ and CC' .

Choose a line through P which is not in E and two points X and Y on it.

The lines XA and YA' lie in a plane, so they intersect in a point \tilde{A} . Thus,

$$\tilde{A} = XA \cap YA',$$

and similarly

$$\tilde{B} = XB \cap YB',$$

$$\tilde{C} = XC \cap YC'.$$

Now A, B, C span E and $\tilde{A}, \tilde{B}, \tilde{C}$ span another plane \tilde{E} . The three lines $A\tilde{A}, B\tilde{B},$ and $C\tilde{C}$ pass through a common point (namely X). Thus, we can apply Theorem 1.1 to the six points $A, \tilde{A}, B, \tilde{B}, C, \tilde{C}$, and find that the line of intersection $E \cap \tilde{E}$ contains

$$A'' = BC \cap \tilde{B}\tilde{C} = BC \cap B'C',$$

and similarly B'' and C'' . \square

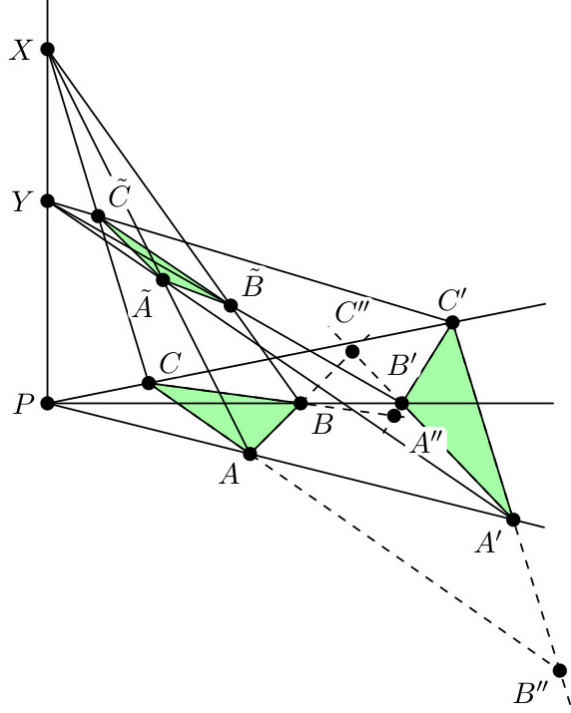


Figure 3. Desargues' theorem in \mathbb{RP}^2 from Desargues' theorem in \mathbb{RP}^3 .

1.7 Duality

As we have seen in Section 1.1, in homogeneous coordinates x_1, x_2, x_3 , the equation for a line in a projective plane is

$$a_1x_1 + a_2x_2 + a_3x_3 = 0,$$

where not all coefficients a_i are zero. The coefficients a_1, a_2, a_3 can be seen as homogeneous coordinates for the line, because if we replace in the equation a_i by λa_i for some $\lambda \neq 0$ we get an equivalent equation for the same line. Thus, the set of lines in a projective plane is itself a projective plane, the *dual plane*. Points in the dual plane correspond to lines in the original plane. Moreover, if we consider in the above equation the x_i as fixed and the a_i as variables, we get an equation for a line in the dual plane. Points on this line correspond to lines in the original plane that contain $[x]$. Thus, a the points on a line in the dual plane correspond to lines in the original plane through a point.

It makes sense to look at this phenomenon in a basis independent way and for arbitrary dimension. It boils down to the duality of vector spaces.

1.7.1 Dual space

The *dual vector space* of \mathbb{R}^{n+1} is the space of linear functionals $\mathbb{R}^{n+1} \rightarrow \mathbb{R}$

$$(\mathbb{R}^{n+1})^* := \{a \mid a : \mathbb{R}^{n+1} \rightarrow \mathbb{R} \text{ linear}\}.$$

The *dual projective space* of \mathbb{RP}^n is correspondingly defined by

$$(\mathbb{RP}^n)^* := P((\mathbb{R}^{n+1})^*).$$

The natural identification $(\mathbb{R}^{n+1})^{**} = \mathbb{R}^{n+1}$ carries over to the projective setting $(\mathbb{RP}^n)^{**} = \mathbb{RP}^n$.

1.7.2 Dual subspaces

For a projective subspace $P(U) \subset \mathbb{RP}^n$ its *dual projective subspace* $P(U)^* \subset (\mathbb{RP}^n)^*$ is defined by

$$P(U)^* := \{[a] \in (\mathbb{RP}^n)^* \mid a(x) = 0 \text{ for all } x \in U\}.$$

The dimensions of a projective subspace and its dual projective subspace are related by

$$\dim P(U) + \dim P(U)^* = n - 1.$$

Incidences are reversed by duality

$$P(U_1) \subset P(U_2) \quad \Leftrightarrow \quad P(U_2)^* \subset P(U_1)^*.$$

and meet and join are interchanged

$$\begin{aligned} (P(U_1) \vee P(U_2))^* &= P(U_1)^* \cap P(U_2)^*, \\ (P(U_1) \cap P(U_2))^* &= P(U_1)^* \vee P(U_2)^*. \end{aligned}$$

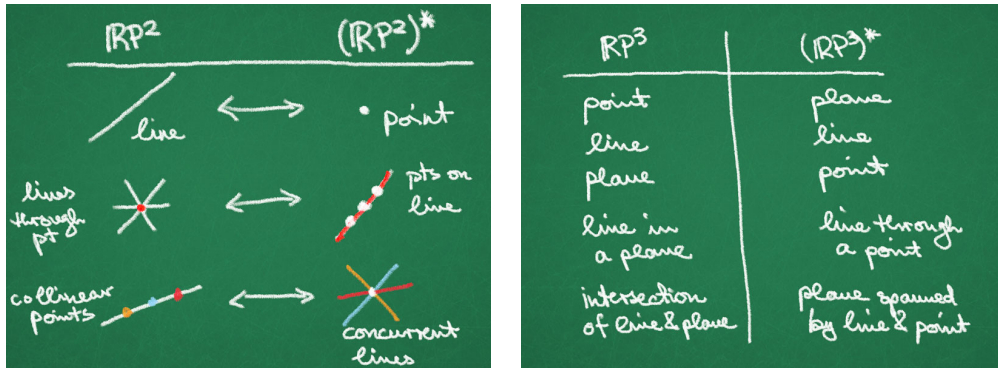


Figure 4. Duality in \mathbb{RP}^2 and \mathbb{RP}^3 .

1.7.3 Duality in coordinates

Let b_1, \dots, b_{n+1} be a basis of \mathbb{R}^{n+1} and b_1^*, \dots, b_{n+1}^* the corresponding dual basis of $(\mathbb{R}^{n+1})^*$, i.e.,

$$b_i^*(b_j) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

In homogeneous coordinates with respect to those bases the duality of two points

$$[x_1, \dots, x_{n+1}] \cong [x] \in \mathbb{RP}^n, \quad [a_1, \dots, a_{n+1}] \cong [a] \in (\mathbb{RP}^n)^*$$

is expressed by

$$a(x) = (a_1 \dots a_{n+1}) \begin{pmatrix} x_1 \\ \vdots \\ x_{n+1} \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_{n+1} \end{pmatrix}^\top \begin{pmatrix} x_1 \\ \vdots \\ x_{n+1} \end{pmatrix} = 0.$$

Thus, duality in linear algebra as well as in projective geometry expresses in a formal way that a subspace can either be expressed as the span of points or the solutions to a set of linear equations.

If a change of basis acts on the homogeneous coordinates of \mathbb{RP}^n as

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{n+1} \end{bmatrix} \mapsto \begin{bmatrix} A \begin{pmatrix} x_1 \\ \vdots \\ x_{n+1} \end{pmatrix} \end{bmatrix}$$

with $A \in \text{GL}(\mathbb{R}^{n+1})$, it acts on the homogeneous coordinates of the dual space $(\mathbb{RP}^n)^*$ as

$$\begin{bmatrix} a_1 \\ \vdots \\ a_{n+1} \end{bmatrix} \mapsto \begin{bmatrix} A^{-\top} \begin{pmatrix} a_1 \\ \vdots \\ a_{n+1} \end{pmatrix} \end{bmatrix}.$$

1.7.4 The dual of Desargues' theorem

The interchangeability of points and lines is called the principle of *duality* in the projective plane. According to this principle, there belongs to every theorem a second theorem that corresponds to it dually, and to every figure a second figure that corresponds to it dually. (Geometry and the Imagination – Hilbert, Cohn-Vossen)

As an example consider the theorem of Desargues in \mathbb{RP}^2 (Theorem 1.2). Then its dual turns out to be the converse statement, which therefore also holds.

1.8 Projective transformations

Let $F \in \text{GL}(\mathbb{R}^{n+1})$ an invertible linear transformation. Then the map

$$[F] : \mathbb{RP}^n \rightarrow \mathbb{RP}^n, \quad [v] \mapsto [F(v)]$$

is called a *projective transformation*.

Proposition 1.3.

(i) *Projective transformations are well-defined maps (do not depend on the representative vectors of points).*

(ii) *For $F, G \in \text{GL}(\mathbb{R}^{n+1})$*

$$[F] = [G] \Leftrightarrow G = \lambda F \text{ with some } \lambda \in \mathbb{R}, \lambda \neq 0.$$

(iii) *Projective transformations map projective subspaces to projective subspaces, while preserving their dimension and incidences.*

(iv) *Vice versa, any bijective map on \mathbb{RP}^n , $n \geq 2$, that maps lines to lines is a projective transformation.*

(v) *Let $A_1, \dots, A_{n+2} \in \mathbb{RP}^n$ be $n+2$ points in general position, and let $B_1, \dots, B_{n+2} \in \mathbb{RP}^n$ be $n+2$ points in general position. Then there exists a unique projective transformation*

$$f : \mathbb{RP}^n \rightarrow \mathbb{RP}^n \quad \text{with} \quad f(A_i) = B_i \text{ for } i = 1, \dots, n+2.$$

(vi) *Projective transformations preserve the cross-ratio of four points on a line.*

1.8.1 Projective transformations in homogeneous coordinates

In homogeneous coordinates a projective transformation $[F] : \mathbb{RP}^n \rightarrow \mathbb{RP}^n$ is represented by a non-singular matrix $F \in \mathbb{R}^{(n+1) \times (n+1)}$ (up to non-zero scalar multiples).

For representative vectors $x = (u_1, \dots, u_n, 1)$ and with

$$F = \left(\begin{array}{c|c} A & b \\ \hline c^\top & d \end{array} \right) \quad \text{where } A \in \mathbb{R}^{n \times n}, b, c \in \mathbb{R}^n, d \in \mathbb{R}$$

we obtain

$$F(x) = \left(\begin{array}{c|c} A & b \\ \hline c^\top & d \end{array} \right) \begin{pmatrix} u \\ 1 \end{pmatrix} = \begin{pmatrix} Au + b \\ c^\top u + d \end{pmatrix} \sim \begin{pmatrix} \frac{Au+b}{c^\top u + d} \\ 1 \end{pmatrix}$$

if $c^\top u + d \neq 0$. Thus, in affine coordinates, projective transformations are *fractional linear transformations*:

$$\mathbb{R}^n \rightarrow \mathbb{R}^n \quad u \mapsto \frac{Au+b}{c^\top u + d}$$

1.8.2 Affine transformations

If we choose a representative matrix of the form

$$F = \left(\begin{array}{c|c} A & b \\ \hline 0 & 1 \end{array} \right) \quad \text{where } A \in \text{GL}(\mathbb{R}^n), b \in \mathbb{R}^n,$$

we obtain

$$\left(\begin{array}{c|c} A & b \\ \hline 0 & 1 \end{array} \right) \begin{pmatrix} u \\ 1 \end{pmatrix} = \begin{pmatrix} Au + b \\ 1 \end{pmatrix}$$

In affine coordinates, this is an *affine transformation*

$$\mathbb{R}^n \rightarrow \mathbb{R}^n \quad u \mapsto Au + b$$

Thus, affine transformations are projective transformations.

Note that affine transformations map the hyperplane at infinity $\{[x] \in \mathbb{RP}^n \mid x_{n+1} = 0\}$ to itself:

$$\left(\begin{array}{c|c} A & b \\ \hline 0 & 1 \end{array} \right) \begin{pmatrix} u \\ 0 \end{pmatrix} = \begin{pmatrix} Au + b \\ 0 \end{pmatrix}$$

In fact, affine transformations are characterized by this property among the projective transformations.

Proposition 1.4. *A projective transformation $f : \mathbb{RP}^n \rightarrow \mathbb{RP}^n$ is an affine transformation if and only if f maps the hyperplane at infinity $\{[x] \in \mathbb{RP}^n \mid x_{n+1} = 0\}$ to itself.*

1.8.3 Euclidean transformations

Euclidean transformations are affine transformations, and thus, projective transformations. Indeed, if we choose a representative matrix of the form

$$F = \left(\begin{array}{c|c} A & b \\ \hline 0 & 1 \end{array} \right) \quad \text{where } A \in \text{O}(n), b \in \mathbb{R}^n,$$

in affine coordinates, this is a Euclidean transformation.

Example 1.8 (reflection in a line). Consider a line with unit normal $n = (n_1, n_2) \in \mathbb{S}^1$ through the point $q \in \mathbb{R}^2$

$$\ell = \{u = (u_1, u_2) \in \mathbb{R}^2 \mid \langle n, u - q \rangle = 0\}$$

Then the (Euclidean) reflection $\hat{\sigma} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is given by

$$\hat{\sigma}(u) = u - 2\langle u - q, n \rangle n$$

With $h := -\langle q, n \rangle$ the equation for the line becomes

$$\langle n, u \rangle + h = 0$$

and the reflection can be rewritten as

$$\hat{\sigma}(u) = u - 2\langle u, n \rangle n - 2hn = (I - 2nn^\top)u - 2hn$$

Thus, in homogeneous coordinates we can write

$$\begin{pmatrix} \hat{\sigma}(u) \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} I - 2nn^\top & -2hn \\ 0 & 1 \end{pmatrix}}_{=: F} \begin{pmatrix} u \\ 1 \end{pmatrix},$$

where, indeed, $I - 2nn^\top \in \text{O}(2)$. As an extension of $\hat{\sigma}$, we can now define a projective transformation $\sigma : \mathbb{RP}^2 \rightarrow \mathbb{RP}^2$ by $\sigma([x]) = [Fx]$. Note that $F^2 = I$ and thus σ is an involution: $\sigma \circ \sigma = \text{id}$.

Let us also derive the matrix F for the reflection in the case that the line is given in homogeneous coordinates

$$\ell = \{[x] \in \mathbb{RP}^2 \mid a^\top x = a_1x_1 + a_2x_2 + a_3x_3 = 0\} = [a]^\star \quad \text{with some } a \in \mathbb{R}^3 \setminus \{0\}$$

With $\hat{a} := (a_1, a_2)$ and $|\hat{a}| \neq 0$ it relates to the Euclidean equation by

$$n = \frac{\hat{a}}{|\hat{a}|}, \quad h = \frac{a_3}{|\hat{a}|}.$$

Thus,

$$F = \left(\begin{array}{c|c} I - 2\frac{\hat{a}\hat{a}^\top}{|\hat{a}|^2} & -2\frac{a_3\hat{a}}{|\hat{a}|^2} \\ \hline 0 & 1 \end{array} \right) \sim \left(\begin{array}{c|c} |\hat{a}|^2 I - 2\hat{a}\hat{a}^\top & -2a_3\hat{a} \\ \hline 0 & |\hat{a}|^2 \end{array} \right)$$

Note that this formula easily generalizes to the (Euclidean) reflection in a hyperplane in $\mathbb{R}^n \subset \mathbb{RP}^n$ given by

$$L = \{[x] \in \mathbb{RP}^n \mid a^\top x = 0\} = [a]^\star,$$

which yields

$$F = \left(\begin{array}{c|c} |\hat{a}|^2 I - 2\hat{a}\hat{a}^\top & -2a_{n+1}\hat{a} \\ \hline 0 & |\hat{a}|^2 \end{array} \right),$$

where $\hat{a} = (a_1, \dots, a_n)$.

1.8.4 Central projections

Another important class of projective transformations are projections.

Example 1.9 (orthogonal projection to a line). Consider a line

$$\ell = \{u = (u_1, u_2) \in \mathbb{R}^2 \mid \langle n, u - q \rangle = \langle n, u \rangle + h = 0\},$$

with some $n \in \mathbb{S}^1$, $q \in \mathbb{R}^2$, and $h = -\langle n, q \rangle$. Then the orthogonal projection $\hat{\pi} : \mathbb{R}^2 \rightarrow \ell$ is given by

$$\hat{\pi}(u) = u - \langle u - q, n \rangle n = u - \langle u, n \rangle n - hn = (I - nn^\top)u - hn$$

Thus, in homogeneous coordinates we can write

$$\begin{pmatrix} \hat{\sigma}(u) \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} I - nn^\top & -hn \\ 0 & 1 \end{pmatrix}}_{=: F} \begin{pmatrix} u \\ 1 \end{pmatrix}.$$

Note that here F is not invertible, since in particular $F \begin{pmatrix} n \\ 0 \end{pmatrix} = 0$. Thus, we can extend $\hat{\pi}$ to a map

$$\pi : \mathbb{RP}^2 \setminus \left\{ \begin{bmatrix} n \\ 0 \end{bmatrix} \right\} \rightarrow \ell$$

by $\pi([x]) = [Fx]$. Since π is not invertible, it does not constitute a projective transformation. But the restriction of π to any line (that does not contain $\begin{bmatrix} n \\ 0 \end{bmatrix}$) is.

Similar, to Example 1.8, this can easily be generalized to the orthogonal projection onto a hyperplane in $\mathbb{R}^n \subset \mathbb{RP}^n$ given by

$$L = \{[x] \in \mathbb{RP}^n \mid a^\top x = 0\} = [a]^\star,$$

which yields

$$F = \left(\begin{array}{c|c} |\hat{a}|^2 I - \hat{a}\hat{a}^\top & -a_{n+1}\hat{a} \\ \hline 0 & |\hat{a}|^2 \end{array} \right), \quad (3)$$

where $\hat{a} = (a_1, \dots, a_n)$.

More generally, let $L \subset \mathbb{RP}^n$ be a hyperplane and $P \in \mathbb{RP}^n$ a point $P \notin L$. Then the *central projection* to L with center P is given by

$$\pi : \mathbb{RP}^n \setminus \{P\} \rightarrow L, \quad X \mapsto (P \vee X) \cap L$$

P and X span a line, since $X \neq P$. This line intersects L in exactly one point, since $P \notin L$. Thus, this map is well-defined.

Let us show that π is indeed given by a linear map on the representative vectors. Let the hyperplane L be given by

$$L = \{[x] \in \mathbb{RP}^n \mid a(x) = 0\} = [a]^\star \quad \text{with some } a \in (\mathbb{R}^{n+1})^* \setminus \{0\}.$$

The image of a point $X = [x] \neq P = [p]$ lies on the line

$$X \vee P = P(\{\lambda x + \mu p \mid \lambda, \mu \in \mathbb{R}^2\}).$$

Thus, the intersection $(X \vee P) \cap L$ is determined by the condition

$$a(\lambda x + \mu p) = \lambda a(x) + \mu a(p) = 0$$

With $\lambda = a(p)$ and $\mu = -a(x)$, we obtain

$$\pi([x]) = [a(p)x - a(x)p],$$

which is indeed linear in x .

Again, this linear map is not invertible, since p is in its kernel. Furthermore, $\dim \mathbb{RP}^n = n > \dim L = n - 1$. Yet the map becomes a projective transformation once we restrict it to another hyperplane K with $P \notin K$:

$$\pi : K \rightarrow L \quad X = [x] \mapsto (P \vee X) \cap L = [a(p)x - a(x)p]$$

To see that now it is invertible, note that $\dim K = \dim L$. Further $a(p)x - a(x)p = 0$ implies $x = 0$, otherwise we would have $[x] = [p]$, which contradicts $P \notin K$.

In homogeneous coordinates, we can write the representative matrix for the central projection as

$$F = a^\top p I - p a^\top.$$

Example 1.10 (orthogonal projection as central projection). Let us recover the orthogonal projection from Example 1.9 as central projection with center at infinity.

Consider the hyperplane

$$L = \{[x] \in \mathbb{RP}^n \mid a^\top x = 0\} = [a]^\star \quad \text{with some } a \in (\mathbb{R}^{n+1})^* \setminus \{0\}.$$

and $P = [p] = [\hat{a}, 0] = [a_1, \dots, a_n, 0]$. Then

$$\begin{aligned} F &= a^\top p I - p a^\top \\ &= \begin{pmatrix} \hat{a}^\top & a_{n+1} \end{pmatrix} \begin{pmatrix} \hat{a} \\ 0 \end{pmatrix} I - \begin{pmatrix} \hat{a} \\ 0 \end{pmatrix} \begin{pmatrix} \hat{a}^\top & a_{n+1} \end{pmatrix} \\ &= |\hat{a}|^2 I - \left(\begin{array}{c|c} \hat{a}\hat{a}^\top & a_{n+1}\hat{a} \\ \hline 0 & 0 \end{array} \right), \end{aligned}$$

which indeed coincides with (3).

The definition for central projections can be generalized further by decreasing the dimension of the image space which at the same time increasing the dimension of the center.

Let $L, C \subset \mathbb{RP}^n$ be projective subspaces with

$$C \cap L = \emptyset, \quad C \vee L = \mathbb{RP}^n$$

Then the map

$$\pi : \mathbb{RP}^n \setminus C \rightarrow L, \quad X \mapsto (C \vee X) \cap L$$

is called (*generalized*) *central projection* onto L with center C . Indeed, this map is well-defined, since $\dim(C \vee X) = \dim C + 1$ and $\dim L + \dim C = n - 1$ and therefore, $C \vee X$ and L intersect in exactly one point.

Again, the map π becomes invertible and in particular a projective transformation,

$$\pi : K \rightarrow L, \quad X \mapsto (C \vee X) \cap L$$

once restricted to any subspace $K \subset \mathbb{RP}^n$ with

$$\dim K = \dim L, \quad C \cap K = \emptyset$$

Example 1.11 (central projection). If L is a hyperplanes, i.e. $\dim L = n - 1$, the center C is a point, and the generalized central projection becomes the standard central projection.

Example 1.12 (three skew lines). If $n = 3$ and K, L are two non-intersecting lines, then the center C is another line, and we obtain three skew lines.

2 Plane curves and envelopes of lines

2.1 Plane curves

Definition 2.1.

- (i) A (*plane*) *curve* is a smooth map

$$\gamma : I \rightarrow \mathbb{R}^2$$

with some interval $I \subset \mathbb{R}$.

- (ii) Let γ be a curve.

- The vectors

$$\dot{\gamma}(t)$$

are called the *velocity* or *tangent vectors* of γ .

- The function

$$v(t) := \|\dot{\gamma}(t)\|$$

is called the *speed* of γ .

- The function

$$s(t) := \int_{t_1}^t v(t) dt$$

is called the *arc-length* of γ , here $I = [t_1, t_2]$.

- If $v(t) = 1$ for all $t \in I$, then γ is called *arc-length parametrized*.

- (iii) A curve γ is called *regular* if

$$\dot{\gamma}(t) \neq 0 \quad \text{for all } t \in I.$$

- (iv) Let γ be a regular curve and $t \in I$.

- Any non-zero scalar multiple of $\dot{\gamma}(t)$ is called a *tangent vector* at $t \in I$.

- The line

$$T(t) := \{\gamma(t) + \alpha \dot{\gamma}(t) \mid \alpha \in \mathbb{R}\}$$

is called the *tangent line* at $t \in I$.

- Any vector $n(t)$ orthogonal to $\dot{\gamma}(t)$, i.e.,

$$\langle n(t), \dot{\gamma}(t) \rangle = 0,$$

is called a *normal vector* at $t \in I$. In particular one can choose.

$$n(t) = \frac{1}{v(t)} J \dot{\gamma}(t), \quad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

which is called the *unit normal vector* at $t \in I$.

- The line

$$\begin{aligned} N(t) &:= \{x \in \mathbb{R}^2 \mid \langle \dot{\gamma}(t), x - \gamma(t) \rangle = 0\} \\ &= \{\gamma(t) + \alpha n(t) \mid \alpha \in \mathbb{R}\}. \end{aligned}$$

is called the *normal line* at $t \in I$.

Note that the derivative of the arc-length is the speed

$$\dot{s}(t) = v(t).$$

For a regular curve γ the arc-length $s(\cdot)$ is monotonically increasing, and thus invertible. We call its inverse function $t(\cdot) = s^{-1}(\cdot)$ and thus write

$$\gamma(s) = (\gamma \circ t)(s).$$

For the derivative w.r.t. arc-length we write

$$\gamma' = \frac{d}{ds}\gamma = \frac{dt}{ds} \frac{d}{dt}\gamma = \frac{1}{v}\dot{\gamma}.$$

In particular, the parametrization of γ w.r.t. arc-length has unit speed

$$\|\gamma'\| = 1,$$

which implies

$$0 = \frac{d}{ds} \|\gamma'\|^2 = \frac{d}{ds} \langle \gamma', \gamma' \rangle = 2 \langle \gamma'', \gamma' \rangle.$$

Thus γ'' always points in normal direction.

Definition 2.2. Let γ be a regular curve, and let n be the unit normal vector field of γ . Then

$$\kappa(s) = \langle \gamma''(s), n(s) \rangle$$

is called the (*signed*) *curvature* of γ at s , i.e.

$$\gamma''(s) = \kappa(s)n(s).$$

In terms of an arbitrary parametrization, and with unit tangent vector

$$\tau(t) := \frac{\dot{\gamma}(t)}{v(t)}$$

the curvature can be written as

$$\begin{aligned} \kappa(t) &= \frac{1}{v(t)} \langle \dot{\tau}(t), n(t) \rangle = \frac{1}{v(t)^2} \langle \ddot{\gamma}(t), n(t) \rangle \\ &= \frac{1}{v(t)^3} \langle \ddot{\gamma}(t), J\dot{\gamma}(t) \rangle = \frac{1}{v(t)^3} \det(\dot{\gamma}(t), \ddot{\gamma}(t)). \end{aligned}$$

Example 2.1. Consider a parametrized circle of radius $r > 0$

$$\gamma(t) = r \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}, \quad t \in [0, 2\pi].$$

Then

$$\begin{aligned} \dot{\gamma}(t) &= r \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix}, \quad v(t) = \|\dot{\gamma}(t)\| = r, \quad \tau(t) = \frac{\dot{\gamma}(t)}{v(t)} = \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix}, \\ n(t) &= J\tau(t) = \begin{pmatrix} -\cos(t) \\ -\sin(t) \end{pmatrix}, \quad \dot{\tau}(t) = \begin{pmatrix} -\cos(t) \\ -\sin(t) \end{pmatrix}. \end{aligned}$$

Thus, the curvature of γ is

$$\kappa(t) = \frac{1}{v(t)} \langle \dot{\tau}(t), n(t) \rangle = \frac{1}{r}.$$

Definition 2.3. Let $\gamma : I \rightarrow \mathbb{R}^2$ be a regular curve, and n its unit normal vector field. If $\kappa(t) \neq 0$, then the *osculating circle* at $t \in I$ is the circle with center

$$c(t) = \gamma(t) + \frac{1}{\kappa(t)}n(t)$$

and radius

$$r(t) = \frac{1}{|\kappa(t)|}.$$

If $\kappa(t) = 0$, then we consider the tangent line at $t \in I$ to be the osculating circle.

The osculating circle touches its curve the corresponding point. Furthermore, if parametrized in the same direction as the curve, it has the same (signed) curvature.

It can also be shown that it is the best approximating circle in the following sense. Consider the circle through three points of the curve $\gamma(t)$, $\gamma(t - \epsilon)$, and $\gamma(t + \epsilon)$. Then in the limit $\epsilon \rightarrow 0$, this circle converges to the osculating circle.

2.2 Discrete plane curves

Definition 2.4.

- (i) A *discrete (plane) curve* is a map

$$\gamma : I \rightarrow \mathbb{R}^2$$

with some interval $I \subset \mathbb{Z}$. We denote its *vertices* by

$$\gamma_k = \gamma(k) \quad \text{for } k \in I.$$

- (ii) Let γ be a discrete curve.

- The vectors

$$\Delta\gamma_k := \gamma_{k+1} - \gamma_k$$

are called *discrete velocity vectors*, *vertex difference vectors*, or *edge tangent vectors*. They are naturally defined on edges $(k, k + 1)$.

- We define the *turning angle* at a vertex $k \in I$ by

$$\varphi_k := \angle(\Delta\gamma_k, \Delta\gamma_{k-1}) \in [-\pi, \pi].$$

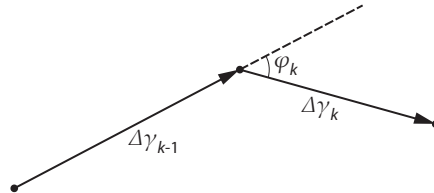


Figure 5. Turning angle at a vertex of a discrete curve.

- If

$$\|\Delta\gamma_k\| = \|\gamma_{k+1} - \gamma_k\| = 1$$

then γ is called *discrete arc-length parametrized curve*.

(iii) A discrete curve γ is called *regular* if any three successive points $\gamma_{k-1}, \gamma_k, \gamma_{k+1}$ are distinct, or equivalently, if any two successive edge tangent vectors are not anti-parallel.

(iv) Let γ be a discrete curve, $k \in I$.

- The line

$$T_k := \gamma_k \vee \gamma_{k+1}$$

is called the *edge tangent line* at the edge $(k, k+1)$.

- The perpendicular bisector N_k of γ_k and γ_{k+1} is called the *edge normal line* at the edge $(k, k+1)$.

We now introduce two types of discrete osculating circles.

Definition 2.5. Let $\gamma : I \rightarrow \mathbb{R}^2$ be a regular discrete curve. Then the circle C_k through three successive points $\gamma_{k-1}, \gamma_k, \gamma_{k+1}$ is called the *vertex osculating circle* at $k \in I$.

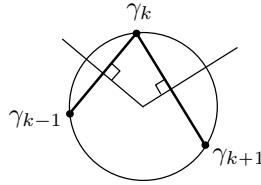


Figure 6. Vertex osculating circle.

- Note that the two involved edge normals N_{k-1} and N_k both contain the center of C_k .
- The discrete curvature at vertex k can now be defined by the radius of the vertex osculating circle. The radius is given by $\|\gamma_{k+1} - \gamma_{k-1}\| = 2R_k \sin \varphi_k$ which leads to the curvature

$$\kappa_k = \frac{2 \sin \varphi_k}{\|\gamma_{k+1} - \gamma_{k-1}\|}.$$

- The vertex osculating circle inherits an orientation from the order of the three points on it. This can be used to also associate a sign to the discrete curvature, which corresponds to the sign in the formula above.
- The vertex osculating circle can also be used to define *vertex tangent lines* as the line tangent to C_k in the point γ_k .

Definition 2.6. Let $\gamma : I \rightarrow \mathbb{R}^2$ be a regular discrete curve. Then the circle C_k that touches three consecutive edge tangent lines T_{k-1}, T_k, T_{k+1} is called the *edge osculating circle* at $(k, k+1) \in I$.

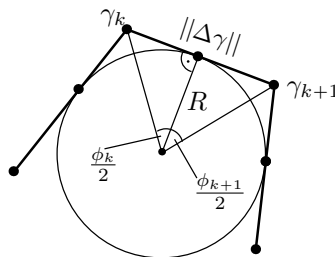


Figure 7. Edge osculating circle.

- For three (non-concurrent) lines in \mathbb{R}^3 there are four circles touching them. By endowing the tangent lines with the orientation coming from the order of the points of the curve on them, this choice can be made unique.

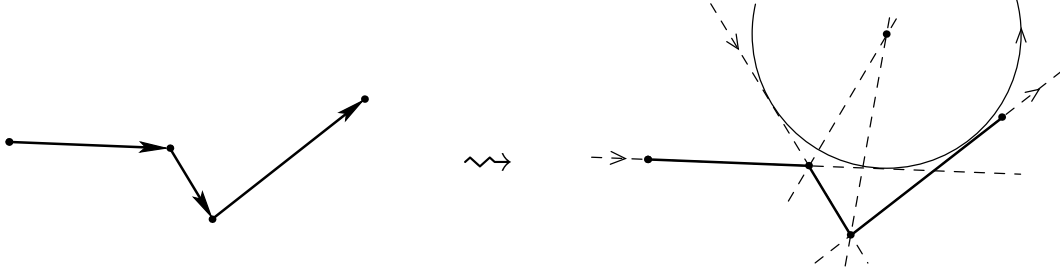


Figure 8. Edge osculating circle from oriented tangent lines.

- Note that the (correctly chosen) angle bisectors of successive edge tangent lines contain the center of the edge osculating circle. Thus, the edge osculating circle can be used to define *edge normal lines*.
- The (oriented) edge osculating circle can be used to define a (signed) discrete curvature at the edge $(k, k + 1)$. The radius is given by $\|\Delta\gamma_k\| = R_k(\tan \frac{\varphi_k}{2} + \tan \frac{\varphi_{k+1}}{2})$. This leads to the curvature

$$\kappa_k = \frac{\tan \frac{\varphi_k}{2} + \tan \frac{\varphi_{k+1}}{2}}{\|\Delta\gamma_k\|}.$$

Computing angle bisectors Consider two oriented lines

$$\ell = \{x \in \mathbb{R}^2 \mid \langle n, x \rangle + h = 0\}, \quad \tilde{\ell} = \{x \in \mathbb{R}^2 \mid \langle \tilde{n}, x \rangle + \tilde{h} = 0\}$$

with $n, \tilde{n} \in \mathbb{S}^1$, $h, \tilde{h} \in \mathbb{R}$, and orientation coming from the normal vectors n, \tilde{n} .

Then the two *angle bisectors* of ℓ and $\tilde{\ell}$ are given by

$$\begin{aligned} m_+ &= \langle x \in \mathbb{R}^2, \langle n + \tilde{n}, x \rangle + h + \tilde{h} = 0 \rangle, \\ m_- &= \langle x \in \mathbb{R}^2, \langle n - \tilde{n}, x \rangle + h - \tilde{h} = 0 \rangle. \end{aligned}$$

Reflection in m_- maps ℓ to $\tilde{\ell}$, but with opposite orientation, while reflection in m_+ maps ℓ to $\tilde{\ell}$ with the same orientation.

Thus, for two adjacent edge tangent lines T_k, T_{k+1} the orientation reversing angle bisector m_- is the desired *vertex normal line*.

2.3 Envelopes

Consider a one-parameter family of curves C (implicitly) given by

$$C(t) = \{x \in \mathbb{R}^2 \mid F(t, x) = 0\}, \quad t \in I$$

with some smooth map $F : I \times \mathbb{R}^2 \rightarrow \mathbb{R}$.

Definition 2.7. A curve $\gamma : I \rightarrow \mathbb{R}^2$ is called *envelope* of the one-parameter family C if γ is tangent to $C(t)$ in the point $\gamma(t)$, i.e.

$$F(t, \gamma(t)) = 0 \quad (\gamma(t) \text{ lies on } C(t)) \quad (4)$$

$$\langle \nabla_x F(t, \gamma(t)), \dot{\gamma}(t) \rangle = 0 \quad (\gamma \text{ in tangent direction of } C(t) \text{ at } \gamma(t)) \quad (5)$$

This is a differential equation for γ . But we can reformulate this in the following way. Equation (4) implies

$$\begin{aligned} 0 &= \frac{d}{dt}F(t, \gamma(t)) = DF(t, \gamma(t)) \begin{pmatrix} 1 \\ \dot{\gamma}(t) \end{pmatrix} = \begin{pmatrix} \partial_t F & \partial_{x_1} F & \partial_{x_2} F \end{pmatrix} \begin{pmatrix} 1 \\ \dot{\gamma}(t) \end{pmatrix} \\ &= \partial_t F(t, \gamma(t)) + \langle \nabla_x F(t, \gamma(t)), \dot{\gamma}(t) \rangle. \end{aligned}$$

Thus, equations (4) and (5) are equivalent to

$$\begin{aligned} F(t, \gamma(t)) &= 0 \\ \partial_t F(t, \gamma(t)) &= 0, \end{aligned}$$

which is not a differential equation in γ anymore.

In particular, if C is a family of lines, then the equations for the envelope are two linear equations in γ .

Example 2.2. For a regular curve $\gamma : I \rightarrow \mathbb{R}^2$ the envelope of its tangent lines is the curve γ itself,

Example 2.3. Consider

$$F(t, x) = x_1 - 2tx_2 + t^2.$$

Then

$$\partial_t F(t, x) = -2x_2 + 2t.$$

implies $x_2 = t$. Substituting this into $F(t, x) = 0$ we obtain $x_1 = t^2$. Thus the envelope is given by

$$\gamma(t) = \begin{pmatrix} t^2 \\ t \end{pmatrix}$$

which is a parabola.

Note, that, in homogeneous coordinates, the equation for the lines is given by

$$x_1 - 2tx_2 + t^2x_3 = \begin{pmatrix} 1 & -2t & t^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$$

which describes a curve $t \mapsto [1, -2t, t^2]$ in $(\mathbb{RP}^2)^*$. This curve is implicitly given by

$$x_2^2 - 4x_1x_3 = 0,$$

which is a conic in $(\mathbb{RP}^2)^*$. This is an example of the general fact, that (the envelope) of the dual of a conic is a conic.

Discrete envelope of a family of lines Let $C : \mathbb{Z} \supset I \rightarrow \text{Lines}(\mathbb{R}^2)$ be a discrete one-parameter family of lines, such that no adjacent lines are equal or parallel.

Then we can define the *discrete envelope* as the discrete curve given by intersections of adjacent lines

$$\gamma_k := C_k \cap C_{k+1}.$$

In this way the edge tangent lines of γ_k coincide with the lines of C ,

$$T_k = C_{k+1}.$$

2.4 Evolute

Definition 2.8. The *evolute* of a regular curve γ is the envelope of its normal lines N .

The envelope of the family of normal lines is described by the equations

$$\begin{aligned} F(t, x) &:= \langle \dot{\gamma}(t), x - \gamma(t) \rangle = 0 \\ \partial_t F(t, x) &= \langle \ddot{\gamma}(t), x - \gamma(t) \rangle - \|\dot{\gamma}(t)\|^2 = 0 \end{aligned}$$

With unit normal field n of γ , the first equation is equivalent to

$$x = e(t) = \gamma(t) + \alpha(t)n(t)$$

with some function α . Then, $\alpha(t)$ can be determined by the second equation

$$\langle \ddot{\gamma}, e(t) - \gamma(t) \rangle - \|\dot{\gamma}(t)\|^2 = \alpha(t) \langle \ddot{\gamma}(t), n(t) \rangle - \|\dot{\gamma}(t)\|^2 = 0$$

to be

$$\alpha(t) = \frac{\|\dot{\gamma}\|^2}{\langle \ddot{\gamma}(t), n(t) \rangle} = \frac{1}{\kappa(t)},$$

which is well-defined as long as $\langle \ddot{\gamma}(t), n(t) \rangle \neq 0$, i.e., $\kappa(t) \neq 0$. Thus, the evolute of γ is given by

$$e(t) = \gamma(t) + \frac{1}{\kappa(t)}n(t),$$

and we find

Proposition 2.1. *The evolute of a regular curve consists of the centers of its osculating circles.*

Proposition 2.2. *Let $\gamma : I \rightarrow \mathbb{R}^2$ be a regular curve. Then its evolute e is non-regular in $t \in I$ if and only if the curvature κ of γ has a local extremum in $t \in I$, i.e.,*

$$\dot{e}(t) = 0 \quad \Leftrightarrow \quad \dot{\kappa}(t) = 0$$

Proof. Let γ be arc-length parametrized. Then

$$e'(s) = \gamma'(s) + \left(\frac{1}{\kappa(s)} \right)' n(s) + \frac{1}{\kappa(s)} n'(s).$$

For the normal vector we have $0 = \frac{d}{ds} \langle n(s), n(s) \rangle = 2 \langle n'(s), n(s) \rangle$, thus $n'(s) = \alpha(s) \gamma'(s)$ where

$$\alpha(s) = \langle n'(s), \gamma'(s) \rangle = - \langle n(s), \gamma''(s) \rangle = -\kappa(s).$$

So,

$$n'(s) = -\kappa(s) \gamma'(s)$$

Thus,

$$e'(s) = \left(\frac{1}{\kappa(s)} \right)' n(s) = -\frac{\kappa'(s)}{\kappa(s)^2} n(s)$$

□

Definition 2.9. A *parallel curve* of γ is a curve of the form

$$\gamma_r(t) := \gamma(t) + rn(t), \quad r \in \mathbb{R}.$$

where n is the unit normal vector field of γ

Proposition 2.3. *Parallel curves have the same evolutes.*

Proof. We show that parallel curves have the same normal lines.

$$\langle \dot{\gamma}_r(t), n(t) \rangle = \langle \dot{\gamma}(t) + r\dot{n}(t), n(t) \rangle = 0.$$

□

Example 2.4. Consider a parabola

$$\gamma(t) := \begin{pmatrix} t \\ t^2 \end{pmatrix}$$

Then

$$\dot{\gamma}(t) = \begin{pmatrix} 1 \\ 2t \end{pmatrix}, \quad \ddot{\gamma}(t) = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \quad n(t) = J\dot{\gamma}(t) \begin{pmatrix} -2t \\ 1 \end{pmatrix}$$

and

$$\langle \ddot{\gamma}(t), n(t) \rangle = 2, \quad \|\dot{\gamma}(t)\|^2 = 1 + 4t^2.$$

Therefore, the evolute is given by

$$e(t) = \gamma(t) + \frac{\|\dot{\gamma}\|^2}{\langle \ddot{\gamma}(t), n(t) \rangle} n(t) = \begin{pmatrix} -4t^3 \\ \frac{1}{2} + 3t^2 \end{pmatrix},$$

which is a semicubic parabola.

Note that it has a cusp at the point where the parabola has maximal curvature.

Discrete evolutes Let $\gamma : \mathbb{Z} \supset I \rightarrow \mathbb{R}^2$ be a regular discrete curve.

- We can define its *vertex evolute* as the discrete envelope of adjacent edge normal lines. The vertex evolute consists of the centers of the vertex osculating circles.
- Alternatively, we can define its *edge evolute* as the discrete envelope of adjacent vertex normal lines. The edge evolute consists of the centers of the edge osculating circles.

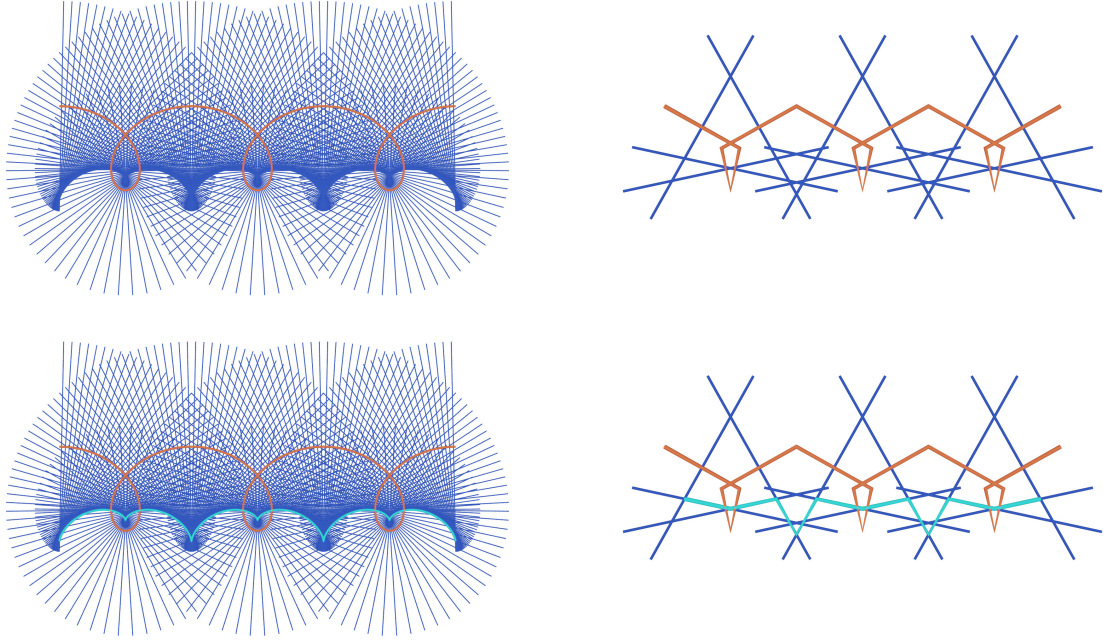


Figure 9. Top: Smooth and discrete curve and its tangent lines. Bottom: Smooth and discrete curve and its evolute.

2.5 Involute

Definition 2.10. An *involute* of a regular curve γ is a curve orthogonal to the tangent lines.

Thus, an involute $\Gamma : I \rightarrow \mathbb{R}^2$ must satisfy

$$\Gamma(t) = \gamma(t) + \alpha(t)\tau(t), \quad \tau(t) = \frac{\dot{\gamma}(t)}{\|\dot{\gamma}(t)\|}$$

with some $\alpha : I \rightarrow \mathbb{R}$ and

$$0 = \langle \dot{\Gamma}(t), \dot{\gamma}(t) \rangle = \langle \dot{\gamma}(t), \dot{\gamma}(t) + \dot{\alpha}(t)\tau(t) + \alpha(t)\dot{\tau}(t) \rangle = \|\dot{\gamma}(t)\|^2 + \dot{\alpha}(t) \|\dot{\gamma}\|.$$

Thus,

$$\dot{\alpha}(t) = -\|\dot{\gamma}(t)\|$$

We obtain

$$\Gamma_a(t) = \gamma(t) - \frac{\dot{\gamma}(t)}{\|\dot{\gamma}(t)\|} \int_a^t \|\dot{\gamma}(t)\| dt = \gamma(t) - \frac{\dot{\gamma}(t)}{\|\dot{\gamma}(t)\|} (s(t) - s(a)),$$

where s is the arc-length of γ .

Thus, in terms of arc-length parametrization the involute is given by

$$\Gamma_a(s) = \gamma(s) - \gamma'(s)(s - a).$$

The distance of the involute to the corresponding curve (along the tangent line) satisfies

$$\|\Gamma_a(s) - \gamma(s)\| = |s - a|.$$

- Thus, the involute is the locus of a point on a piece of taut string as the string is either unwrapped from or wrapped around the curve starting at the point $\gamma(a)$.
- Equivalently, it is the locus of the point on a straight line as it rolls without slipping along the curve.

Proposition 2.4. *Let γ be a regular curve.*

- (i) *The involute is regular at points where $\kappa(t) \neq 0$ and $t \neq a$.*
- (ii) *The normal lines of the involute are the tangents of γ .*
- (iii) *The evolute of the involute is γ .*
- (iv) *The involutes are parallel curves.*

Proof.

- (i) $\Gamma'_a(s) = \gamma'(s) - \gamma''(s)(s-a) - \gamma'(s) = -(s-a)\kappa(s)n(s)$.
- (ii) By definition of the involute $\langle \Gamma'_a(s), \gamma'(s) \rangle = 0$.
- (iii) Follows from (ii).
- (iv) $\Gamma_a(s) = \Gamma_0(s) + a\gamma'(s)$, where $\gamma'(s)$ is the unit normal at $\Gamma_0(s)$.

□

Remark 2.1. The one-parameter family of tangent lines of a curve together with its one-parameter family of involutes form an orthogonal coordinate system.

Example 2.5 (Involute of a circle). Consider a parametrized circle of radius $r > 0$

$$\gamma(t) = r \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}, \quad t \in [0, 2\pi].$$

Then

$$\dot{\gamma}(t) = r \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix}, \quad v(t) = \|\dot{\gamma}(t)\| = r, \quad s(t) - s(a) = r(t - a).$$

Thus, the involutes of γ are given by

$$\Gamma_a(t) = r \begin{pmatrix} \cos(t) - (t-a)\sin(t) \\ \sin(t) + (t-a)\cos(t) \end{pmatrix}$$

This is a common shape for the teeth of gears, the so called “involute gears”.

Example 2.6 (Involute of a semi-cubic). Consider the semicubic parabola, we obtained as the evolute of a parabola. We reconstruct the parabola as one involute of semicubic parabola.

$$\gamma(t) = \begin{pmatrix} -4t^3 \\ \frac{1}{2} + 3t^2 \end{pmatrix}, \quad t > 0.$$

Then

$$\dot{\gamma}(t) = \begin{pmatrix} -12t^2 \\ 6t \end{pmatrix}, \quad \|\dot{\gamma}(t)\| = 6t\sqrt{1+4t^2}, \quad \int_0^t \|\dot{\gamma}(t)\| dt = \frac{1}{2}(1+4t^2)^{\frac{3}{2}} - \frac{1}{2}.$$

For simplicity, we add a constant of integration $\frac{1}{2}$ and obtain

$$\Gamma_0(t) = \begin{pmatrix} -4t^3 \\ \frac{1}{2} + 3t^2 \end{pmatrix} + \frac{1}{6t\sqrt{1+4t^2}} \begin{pmatrix} -12t^2 \\ 6t \end{pmatrix} \frac{1}{2}(1+4t^2)^{\frac{3}{2}} = \begin{pmatrix} t \\ t^2 \end{pmatrix},$$

which is a parabola.

Note that the other involutes of the semicubic parabola are not parabolas.

Discrete involutes We can derive constructions for discrete involutes from the property that evolve of the involute should be the original curve, i.e., the tangent lines of the original curve should be the normal lines of the evolute.

Let $\gamma : \mathbb{Z} \subset I \rightarrow$ be a regular discrete curve.

- Choose some starting point $\Gamma_0 \in \mathbb{R}^2$
- Obtain Γ_{k+1} from Γ_k by reflection in tangent line T_k of γ .

Then T_k is the edge normal line of Γ at the edge $(k, k+1)$.

Alternatively:

- Choose some starting edge tangent line $\tilde{T}_0 = \Gamma_0 \vee \Gamma_1$.
- Obtain \tilde{T}_{k+1} from \tilde{T}_k by reflection in tangent line T_k of γ , and thus, $\Gamma_{k+1} = \tilde{T}_k \cap T_{k+1}$.

Then T_k is the vertex normal line of Γ at the vertex $k+1$.

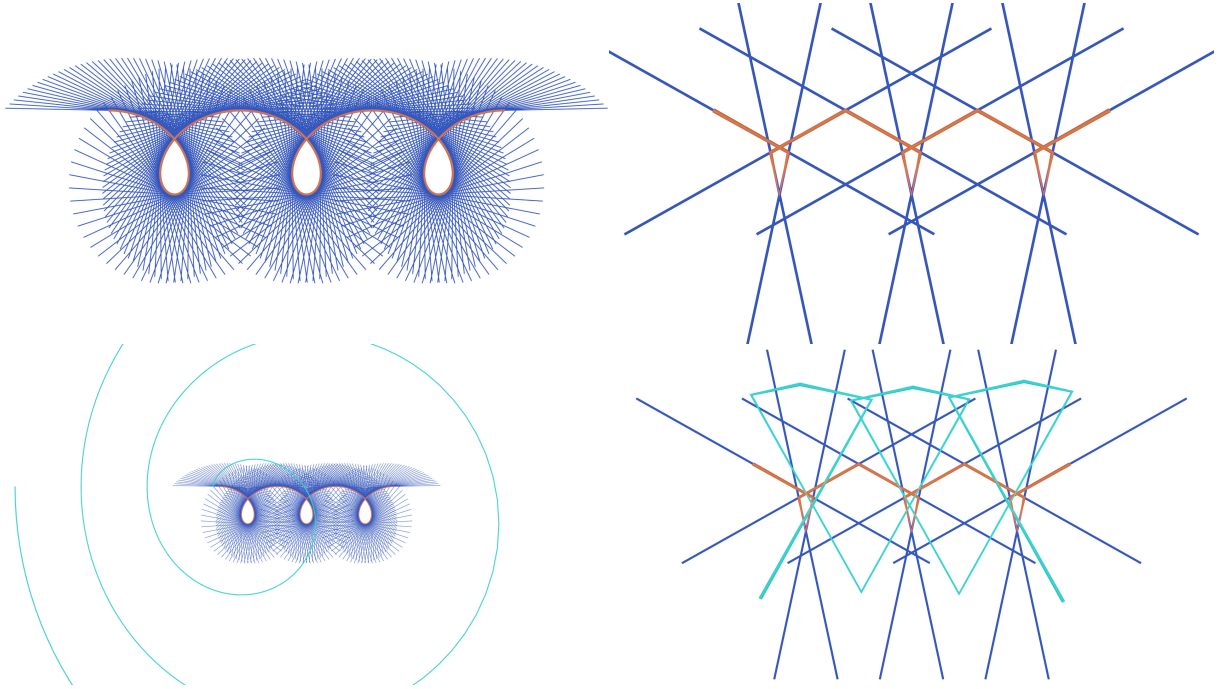


Figure 10. Top: Smooth and discrete curve and its normal lines. Bottom: Smooth and discrete curve and one of its involutes.

3 Conics and quadrics

While projective subspaces are described by linear homogeneous equations, we now add the objects that are described by quadratic homogeneous equations.

Conics or *conic sections* are planar sections of a cone of revolution (or a cylinder)

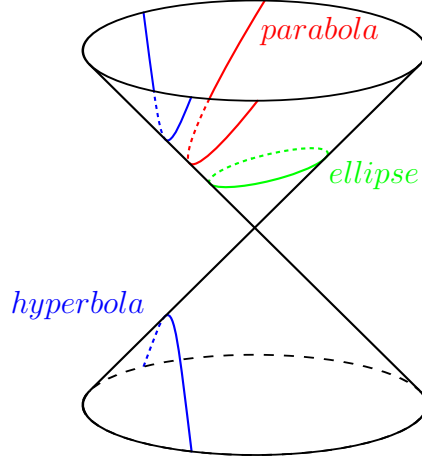


Figure 11. Ellipse, hyperbola, and parabola as a planar section of a cone.

It can be shown that conic sections correspond exactly to the sets of solutions of quadratic equations

$$\{(x, y) \in \mathbb{R}^2 \mid q_{11}x^2 + 2q_{12}xy + q_{22}y^2 + 2q_{13}x + 2q_{23}y + q_{33} = 0.\}$$

Introducing homogeneous coordinates $x = \frac{x_1}{x_3}$, $y = \frac{x_2}{x_3}$, the (non-homogeneous) quadratic equation in 2 variables can be written as a homogeneous quadratic equation in 3 variables

$$q_{11}x_1^2 + 2q_{12}x_1x_2 + q_{22}x_2^2 + 2q_{13}x_1x_3 + 2q_{23}x_2x_3 + q_{33}x_3^2 = 0,$$

or equivalently,

$$b(x, x) := \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \underbrace{\begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{12} & q_{22} & q_{23} \\ q_{13} & q_{23} & q_{33} \end{pmatrix}}_{=:Q} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$$

where Q is a symmetric matrix, i.e. $Q^T = Q$, and b is a symmetric bilinear form on \mathbb{R}^3

$$b : \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}.$$

Example 3.1. An *ellipse* is a conic section. In normal form in \mathbb{R}^2 (up to a Euclidean transformation) it is given by

$$\left\{ (x, y) \in \mathbb{R}^2 \mid \left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1 \right\}.$$

Introducing homogeneous coordinates $x = \frac{x_1}{x_3}$, $y = \frac{x_2}{x_3}$, we can write its equation as a homogeneous quadratic equation

$$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} - x_3^2 = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} \frac{1}{a^2} & & \\ & \frac{1}{b^2} & \\ & & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0.$$

3.1 Bilinear forms

Let V be a vector space over \mathbb{R} of dimension $n + 1$.

A *bilinear form* on V is a map

$$b : V \times V \rightarrow \mathbb{R}$$

which is linear in both arguments.

Let e_1, \dots, e_{n+1} be a basis of V . Then the matrix $Q = (q_{ij}) \in \mathbb{R}^{(n+1) \times (n+1)}$

$$q_{ij} := b(e_i, e_j) \quad \text{for } i, j = 1, \dots, n + 1$$

is called the *representative matrix*, or *Gram matrix*, of the bilinear form b .

For two coordinate vectors $x = \sum_i x_i e_i, y = \sum_i y_i e_i \in V$ we have

$$b(x, y) = x^\top Q y.$$

A change of coordinates $\tilde{x} = Ax$ with $A \in \text{GL}(n + 1)$ acts on the representative matrix as

$$\tilde{Q} = A^{-\top} Q A^{-1}.$$

Symmetric bilinear forms and quadratic forms

A bilinear form is called *symmetric* if

$$b(x, y) = b(y, x) \quad \text{for } x, y \in V,$$

or equivalently, if its representative matrix is symmetric

$$Q^\top = Q.$$

The space of symmetric bilinear forms $\text{Sym}(V)$ is a linear subspace of dimension

$$\dim \text{Sym}(V) = \frac{(n + 1)(n + 2)}{2}.$$

A symmetric bilinear form $b(\cdot, \cdot)$ defines a corresponding *quadratic form* $b(\cdot)$

$$b(x) := b(x, x) \quad \text{for } x \in V.$$

Vice versa, a quadratic form uniquely determines its bilinear form (polarization identity)

$$2b(x, y) = b(x + y) - b(x) - b(y),$$

and thus, the vector spaces of symmetric bilinear forms on V and quadratic forms on V are isomorphic.

3.2 Quadrics

Definition 3.1. Let V be a vector space over \mathbb{R} of dimension $n + 1$, and b a non-zero symmetric bilinear form on V . Then the zero set

$$\mathcal{Q}_b := \{[x] \in \text{P}(V) \mid b(x) = 0\}.$$

is called a quadric in $\text{P}(V)$.

Example 3.2. The quadratic form

$$b(x) = x_1^2 + x_2^2 - x_3^2$$

defines a quadric (conic) in \mathbb{RP}^2

$$\mathcal{Q}_b = \{[x] \in \mathbb{RP}^2 \mid b(x) = x_1^2 + x_2^2 - x_3^2 = 0\}$$

In affine coordinates $x_3 = 1$ this is a circle

$$x_1^2 + x_2^2 = 1.$$

A non-zero scalar multiple of b defines the same quadric:

$$\mathcal{Q}_b = \mathcal{Q}_{\lambda b} \quad \text{for } \lambda \neq 0.$$

Remark 3.1. For some very degenerate images, e.g. if \mathcal{Q}_b is empty, the reverse statement is not true over \mathbb{R} . However, if we either exclude these cases, or consider the complexification of real quadrics, it holds that

$$\mathcal{Q}_b^{\mathbb{C}} = \mathcal{Q}_{\tilde{b}}^{\mathbb{C}} \quad \Leftrightarrow \quad b = \lambda \tilde{b} \quad \text{for some } \lambda \neq 0.$$

Example 3.3. The quadratic forms

$$b(x) = x_1^2 + x_2^2 + x_3^2, \quad \tilde{b}(x) = x_1^2 + 4x_2^2 + x_3^2,$$

both define empty conics in \mathbb{RP}^2

$$\mathcal{Q}_b = \mathcal{Q}_{\tilde{b}} = \emptyset$$

even though $b \neq \lambda \tilde{b}$ for all $\lambda \neq 0$. However, the point $[1, i, 0]$ is contained in $\mathcal{Q}_b^{\mathbb{C}}$, but not in $\mathcal{Q}_{\tilde{b}}^{\mathbb{C}}$. Thus,

$$\mathcal{Q}_b^{\mathbb{C}} \neq \mathcal{Q}_{\tilde{b}}^{\mathbb{C}}.$$

Thus, we can identify the space of quadrics with the projective space $\mathbf{P} \operatorname{Sym}(V)$. Its dimension is given by

$$\dim \mathbf{P} \operatorname{Sym}(V) = \dim \operatorname{Sym}(V) - 1 = \frac{(n+1)(n+2)}{2} - 1 = \frac{n(n+3)}{2}.$$

and the coefficients

$$q_{ij} = b(e_i, e_j), \quad \text{for } j \leq i$$

can be taken as homogeneous coordinates on the space of quadrics.

Determining a quadric through given points

For a point $[x] \in \mathbf{P}(V)$, the quadrics represented by $[Q] \in \mathbf{P} \operatorname{Sym}(V)$ that contain this point are given by the equation

$$x^{\top} Q x = \sum_{i,j=1}^{n+1} x_i x_j q_{ij} = 0,$$

which is an equation in the $\frac{(n+1)(n+2)}{2}$ variables q_{ij} and determines a hyperplane in the space of quadrics $\mathbf{P} \operatorname{Sym}(V)$. Similarly, $\frac{(n+1)(n+2)}{2} - 1$ points determine a system of linear equations, which has a one-dimensional solution space (if all equations are linearly independent). Thus, generically, $\frac{(n+1)(n+2)}{2} - 1$ points uniquely determine a quadric through them.

Example 3.4. In \mathbb{RP}^5 , five points (no four of which are on a line) determine a unique conic.

Singular points of a quadric

A point $[x] \in P(V)$ is called a *singular point* of the quadric \mathcal{Q}_b if $x \in \ker b$, where

$$x \in \ker b = \{x \in V \mid b(x, y) = 0 \text{ for all } y \in V\},$$

or equivalently, in homogeneous coordinates, if

$$Qx = 0.$$

Thus, if $\text{rk } Q = k$ the singular points of \mathcal{Q}_b are contained in a projective subspace of dimension $n - k$, given by

$$P(\ker b).$$

The quadric \mathcal{Q}_b is called *non-degenerate* if it has no singular points, or equivalently, if Q has full rank.

Example 3.5. The quadratic form on \mathbb{R}^3

$$b(x) = x_1^2 - x_2^2 = (x_1 - x_2)(x_1 + x_2)$$

defines a conic in \mathbb{RP}^2 consisting of a pair of lines

$$\begin{aligned} \mathcal{Q}_b &= \{[x] \in \mathbb{RP}^2 \mid b(x) = x_1^2 - x_2^2 = 0\} \\ &= \{[x] \in \mathbb{RP}^2 \mid x_1 - x_2 = 0\} \cup \{[x] \in \mathbb{RP}^2 \mid x_1 + x_2 = 0\} \\ &= \{[\lambda, \pm\lambda, \mu] \in \mathbb{RP}^2 \mid \lambda, \mu \in \mathbb{R}\} \end{aligned}$$

It has one singular point given by $[0, 0, 1]$. Thus it is a degenerate conic.

Lines on a quadric

Lemma 3.1. *If three collinear points are on a conic, then the conic contains the whole line.*

Proof. Exercise. □

3.3 Projective classification of quadrics in \mathbb{RP}^n

Two quadrics $\mathcal{Q}, \tilde{\mathcal{Q}} \subset \mathbb{RP}^n$ are called projectively equivalent if there exists a projective transformation $f : \mathbb{RP}^n \rightarrow \mathbb{RP}^n$ such that

$$f(\mathcal{Q}) = \tilde{\mathcal{Q}}$$

or equivalently, if there exists $F \in \text{GL}(n + 1)$ and $\lambda \in \mathbb{R}$, $\lambda \neq 0$, such that

$$\tilde{Q} = \lambda F^T Q F,$$

where Q and \tilde{Q} are representative matrices for \mathcal{Q} and $\tilde{\mathcal{Q}}$, respectively. Note, that $f = [F^{-1}]$.

By *Sylvester's law of inertia*, there exists an $F \in O(n + 1)$ such that

$$\tilde{Q} = F^T Q F = \text{diag}(\lambda_1, \dots, \lambda_r, \mu_1, \dots, \mu_s, \underbrace{0, \dots, 0}_t)$$

where,

$$\lambda_i > 0, \quad \mu_i < 0, \quad r + s + t = n.$$

Thus, after applying this transformation the equation for the quadric is of the form

$$\lambda_1 x_1^2 + \dots + \lambda_r x_r^2 + \mu_1 x_{r+1}^2 + \dots + \mu_s x_{r+s}^2 = 0$$

By applying a second transformation

$$F = \text{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \dots, \frac{1}{\sqrt{\lambda_r}}, \frac{1}{\sqrt{-\mu_1}}, \dots, \frac{1}{\sqrt{-\mu_s}}, \underbrace{1, \dots, 1}_t\right)$$

we obtain

$$\tilde{Q} = \text{diag}(\underbrace{1, \dots, 1}_r, \underbrace{-1, \dots, -1}_s, \underbrace{0, \dots, 0}_t),$$

or as an equation for the quadric

$$x_1^2 + \dots + x_r^2 - x_{r+1}^2 - \dots - x_{r+s}^2 = 0.$$

The tuple (r, s, t) , also written as

$$(\underbrace{+ \dots +}_r, \underbrace{- \dots -}_s, \underbrace{0 \dots 0}_t),$$

is called the *signature* of the quadric. We define the signature up to the following equivalence

$$(r, s, t) \sim (s, r, t),$$

and obtain the following classification result.

Theorem 3.2. *Two quadrics in \mathbb{RP}^n are projectively equivalent if and only if they have the same signature.*

Quadrics in \mathbb{RP}^1

- ▶ $(++)$ *empty quadric.* By complexification these are two complex conjugate points.
- ▶ $(+-)$ *two points.*
- ▶ $(+0)$ *one (double) point.*

Quadrics in \mathbb{RP}^2 (conics)

- ▶ $(+++)$ *empty conic.* By complexification this is an imaginary conic.
- ▶ $(++-)$ *oval conic.* Its normal form is given by

$$x_1^2 + x_2^2 - x_3^2 = 0$$

In affine coordinates this conic is an ellipse, a hyperbola, or a parabola. Indeed, if we choose $x_3 = 1$, the equation becomes the equation for a circle

$$x_1^2 + x_2^2 = 1.$$

If we choose coordinates $y_1 = x_1, y_2 = x_3, y_3 = x_2$ and $y_3 = 1$, the equation becomes the equation for a hyperbola

$$y_1^2 - y_2^2 = 1$$

If we choose coordinates $y_1 = x_1, y_2 = x_2 + x_3, y_3 = x_3 - x_2$ and $y_3 = 1$, the equation becomes the equation for a parabola

$$y_1^2 = y_2.$$

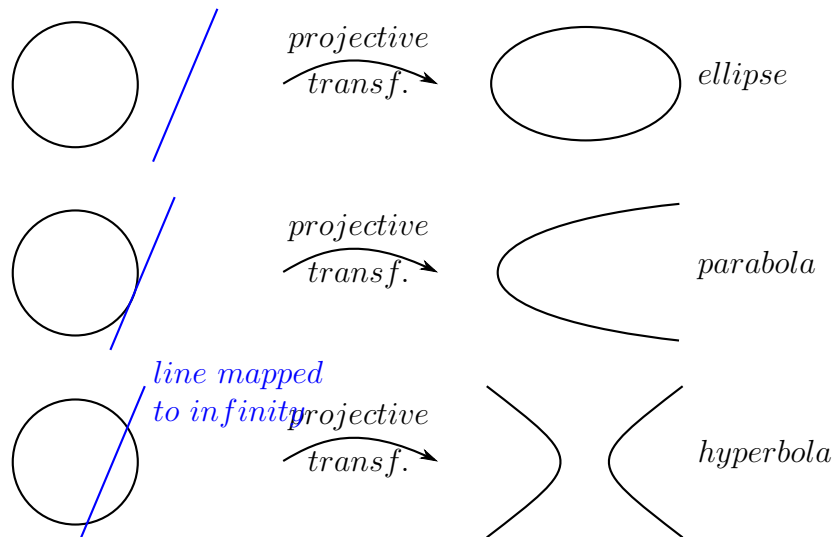


Figure 12. Projective transformations mapping a circle onto an ellipse, a parabola, or a hyperbola.

- $(++0)$ *point*. By complexification these are two imaginary lines that intersect in a real point.
- $(+-0)$ *pair of lines*.
- $(+00)$ *one (double) line*.

Remark 3.2. Note that in \mathbb{RP}^2 only degenerate conics may contain lines.

Quadrics in \mathbb{RP}^3


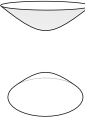



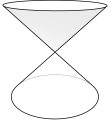

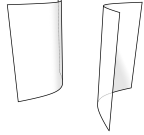
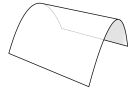

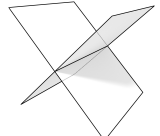
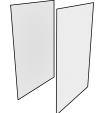



non-degenerate quadrics:			
affine type	affine signature affine normal form	picture	signature projective normal form
ellipsoid	$(+++)_-$ $x^2 + y^2 + z^2 = 1$		$(+++)$ $x_1^2 + x_2^2 + x_3^2 - x_4^2 = 0$
2-sheeted hyperboloid	$(+++)_+$ $x^2 + y^2 - z^2 = -1$		
elliptic paraboloid	$(+++)_p$ $z = x^2 + y^2$		
1-sheeted hyperboloid	$(+ + --)_-$ $x^2 + y^2 - z^2 = 1$		$(+ + --)$ $x_1^2 + x_2^2 - x_3^2 - x_4^2 = 0$
hyperbolic paraboloid	$(+ + --)_p$ $z = x^2 - y^2$		
empty (imaginary)	$(++++)_+$ $x^2 + y^2 + z^2 = -1$		$(++++)$ $x_1^2 + x_2^2 + x_3^2 + x_4^2 = 0$

Table 2. Affine types of non-degenerate quadrics in \mathbb{R}^3 and the corresponding projective types in \mathbb{RP}^3 .

degenerate quadrics:			
affine type	affine signature affine normal form	picture	signature projective normal form
cone	$(++-0)_0$ $x^2 + y^2 - z^2 = 0$		$(++-0)$ $x_1^2 + x_2^2 - x_3^2 = 0$
elliptic cylinder	$(++-0)_-$ $x^2 + y^2 = 1$		
hyperbolic cylinder	$(++-0)_+$ $x^2 - y^2 = 1$		
parabolic cylinder	$(++-0)_p$ $z = x^2$		
one point (imaginary cone)	$(+++0)_0$ $x^2 + y^2 + z^2 = 0$		$(+++0)$ $x_1^2 + x_2^2 + x_3^2 = 0$
empty (imaginary cylinder)	$(+++0)_+$ $x^2 + y^2 = -1$		
two intersecting planes	$(+-00)_0$ $x^2 - z^2 = 0$		$(+-00)$ $x_1^2 - x_2^2 = 0$
two parallel planes	$(+-00)_-$ $x^2 = 1$		
one plane (and one at infinity)	$(+-00)_p$ $x = 0$		
one line (two intersecting imaginary planes)	$(++00)_0$ $x^2 + z^2 = 0$		$(++00)$ $x_1^2 + x_2^2 = 0$
empty (two parallel imaginary planes)	$(++00)_+$ $x^2 = -1$		
one "double" plane	$(+000)_0$ $x^2 = 0$		$(+000)$ $x_1^2 = 0$
empty (one "double" plane at infinity)	$(+000)_+$ $1 = 0$		

34

Table 3. Affine types of degenerate quadrics in \mathbb{R}^3 and the corresponding projective types in \mathbb{RP}^3 .

Remark 3.3. In \mathbb{RP}^3 the non-degenerate quadrics of signature $(++--)$ contain lines.

3.4 Affine classification of quadrics in $\mathbb{R}^n \subset \mathbb{RP}^n$

Two quadrics $\mathcal{Q}, \tilde{\mathcal{Q}} \subset \mathbb{RP}^n$ are called affine equivalent if there exists an affine transformation $f : \mathbb{RP}^n \rightarrow \mathbb{RP}^n$ such that

$$f(\mathcal{Q}) = \tilde{\mathcal{Q}}$$

or equivalently, if there exists $F \in \text{GL}(n+1)$ with

$$F = \left(\begin{array}{c|c} A & b \\ \hline 0 & 1 \end{array} \right), \quad A \in \text{GL}(n), b \in \mathbb{R}^n,$$

and a $\lambda \in \mathbb{R}, \lambda \neq 0$, such that

$$\tilde{Q} = \lambda F^\top Q F.$$

With

$$Q = \left(\begin{array}{c|c} S & q \\ \hline q^\top & \sigma \end{array} \right), \quad S \in \text{Sym}(n), q \in \mathbb{R}^n, \sigma \in \mathbb{R}.$$

we obtain

$$F^\top Q F = \left(\begin{array}{c|c} A^\top S A & A^\top (Sb + q) \\ \hline (b^\top S + q^\top) A & b^\top S b + 2q^\top b + \sigma \end{array} \right),$$

Thus, in a first step, we can use A to bring S to the form

$$S = \text{diag}(\underbrace{1, \dots, 1, -1, \dots, -1}_k, 0, \dots, 0).$$

Case 1: There exists $b \in \mathbb{R}^n$ such that $Sb + q = 0$. Then Q can be brought to the form

$$Q = \left(\begin{array}{c|c} S & 0 \\ \hline 0 & \sigma \end{array} \right), \quad S = \text{diag}(1, \dots, 1, -1, \dots, -1, 0, \dots, 0), \sigma = 0, 1, -1.$$

Here $\sigma = 0, 1, -1$ can be achieved by rescaling Q and then using A to rescale S . If (r, s, t) is the projective signature of \mathcal{Q} , we write the *affine signature* in this case as

$$(r, s, t)_\sigma$$

with

$$(r, s, t)_\sigma \sim (s, r, t)_{-\sigma}$$

Case 2: There exists no $b \in \mathbb{R}^n$ such that $Sb + q = 0$. Then S must be singular, i.e., $k < n$. Now we apply the following steps:

- We choose $b \in \mathbb{R}^n$ such that the first k components of $Sb + q$ vanish.
- We choose A such that $A^\top (Sb + q) = e_n$ without changing S .
- We choose $b = -\frac{\sigma}{2} e_n$ to eliminate σ .

Thus, Q can be brought to the form

$$Q = \left(\begin{array}{c|cc} \hat{S} & & 0 \\ \hline & 0 & 1 \\ 0 & 1 & 0 \end{array} \right), \quad \hat{S} = \text{diag}(1, \dots, 1, -1, \dots, -1, 0, \dots, 0)$$

If (r, s, t) is the projective signature of \mathcal{Q} , we write the *affine signature* in this case as

$$(r, s, t)_p$$

with

$$(r, s, t)_p \sim (s, r, t)_p.$$

Note that the block $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ corresponds to a projective signature of $(+-)$. Thus, an affine signature $(r, s, t)_p$ is only possible with $r > 0$ and $s > 0$.

Theorem 3.3. *Two quadrics in \mathbb{RP}^n are affine equivalent if and only if they have the same affine signature.*

3.5 Signature of subspaces

Let $\mathcal{Q} \subset \mathbb{RP}^n$ be a quadric, and $K = P(U) \subset \mathbb{RP}^n$ a projective subspace. Then the *signature of K* (with respect to \mathcal{Q}) is the signature of \mathcal{Q} restricted to K :

$$\{[x] \in K \mid b(x) = 0\}$$

Thus, it is determined by the restriction of the symmetric bilinear form b to U .

Signature of a point A quadric $\mathcal{Q} \subset \mathbb{RP}^n$ separates \mathbb{RP}^n into two connected components. For point $[x] \in \mathbb{RP}^n$ the signature can take 3 possible values:

- ▶ (+) if $b(x) > 0$. The point lies on one side of \mathcal{Q} .
- ▶ (-) if $b(x) < 0$. The point lies on the other side of \mathcal{Q} .
- ▶ (0) if $b(x) = 0$. The point lies on \mathcal{Q} .

Signature of a line A line $\ell \subset \mathbb{RP}^n$ can have the following possible signatures:

- ▶ (++) The line does not intersect \mathcal{Q} .
- ▶ (+-) The line intersects \mathcal{Q} in two points.
- ▶ (+0) The line intersects \mathcal{Q} in one point.
- ▶ (00) The line is contained in \mathcal{Q} .

If the line is given as the span of two points $\ell = [x] \vee [y]$, the representative matrix for b on the corresponding subspace is given by

$$Q = \begin{pmatrix} b(x, x) & b(x, y) \\ b(x, y) & b(y, y) \end{pmatrix}.$$

Note that its determinant

$$\det Q = b(x, x)b(y, y) - b(x, y)^2$$

is the product of its eigenvalues. Thus, if we exclude the case (00), which corresponds to $Q = 0$, the other three cases can be distinguished by the sign of the determinant. The line ℓ has signature

$$\begin{aligned} (+-) &\Leftrightarrow \det Q < 0, \\ (++) &\Leftrightarrow \det Q > 0, \\ (+0) &\Leftrightarrow \det Q = 0. \end{aligned}$$

3.6 Tangent lines and tangent cones

Let $\mathcal{Q} \subset \mathbb{RP}^n$ be a quadric.

A *tangent line* of \mathcal{Q} is a line that intersects \mathcal{Q} in exactly one point. We have established that these are the lines of signature (+0), and can be characterized in the following way.

Lemma 3.4. *A line $[x] \vee [y]$ not contained in \mathcal{Q} is a tangent line of \mathcal{Q} , if and only if*

$$b(x, x)b(y, y) - b(x, y)^2 = 0.$$

Let $X = [x] \in \mathbb{RP}^n \setminus \mathcal{Q}$ a point not on \mathcal{Q} . Then the *tangent cone* to \mathcal{Q} from P is defined as the union of all tangent lines to \mathcal{Q} that contain the point P :

$$\mathcal{C}_X = \bigcup_{\substack{\ell \ni X, \\ \ell \text{ tangent of } \mathcal{Q}}} \ell = \{[y] \in \mathbb{RP}^n \mid c(y) := b(x, x)b(y, y) - b(x, y)^2 = 0\}.$$

Note that c defines a quadratic form, and thus \mathcal{C}_X is a quadric itself.

By definition, every tangent line has a point on \mathcal{Q} , which we call the *point of tangency*. Thus, to obtain the tangent cone it is sufficient to join X with all points of tangency. By Lemma 3.4, for a point $[y] \in \mathcal{Q}$ on \mathcal{Q} , the line $[x] \vee [y]$ is a tangent line if and only if

$$b(x, y) = 0.$$

Thus, the points of tangency of all tangent lines through X lie in a hyperplane,

$$X^\perp = \{[y] \in \mathbb{RP}^n \mid b(x, y) = 0\}$$

called the *polar hyperplane* of X (with respect to \mathcal{Q}). Thus, we can write the tangent cone in the following way

$$\mathcal{C}_X = \bigcup_{Y \in \mathcal{Q} \cap X^\perp} X \vee Y.$$

Example 3.6 (Shadow of an ellipsoid).

What form does the shadow of an ellipsoid have?

Consider an ellipsoid $\mathcal{E} \subset \mathbb{R}^3 \subset \mathbb{RP}^3$ (an affine image of a sphere). Let X be a point outside \mathcal{E} , and K a plane. The shadow of the ellipsoid cast onto K by a light source in X is bounded by the intersection with (one half of) the tangent cone \mathcal{C}_X . Thus it is a conic section.

Which type of conic section can we obtain? Can it be a hyperbola?

The type of conic section (ellipse, parabola, hyperbola) depends on how many points of intersection (0, 1, 2) it has with the line at infinity on K , or equivalently, how many generators of \mathcal{C}_X intersect K in the line at infinity.

Generally, a line intersects the plane K in the line at infinity, if it is parallel to K . Thus, consider the plane K_X through X parallel to K . Then the number of generators of \mathcal{C}_X in K_K is the number of intersection points of $\mathcal{C}_X \cap K$ with infinity.

Consider the two planes K_1, K_2 parallel to K touching \mathcal{E} in one point. This separates \mathbb{RP}^n into two regions, one containing \mathcal{E} , and one not containing \mathcal{E} .

- ▶ If X is in the region not containing \mathcal{E} , then $\mathcal{C}_X \cap K$ is an ellipse.
- ▶ If X is in the region containing \mathcal{E} , then $\mathcal{C}_X \cap K$ is a hyperbola.
- ▶ If X lies in K_1 or K_2 , then $\mathcal{C}_X \cap K$ is a parabola.

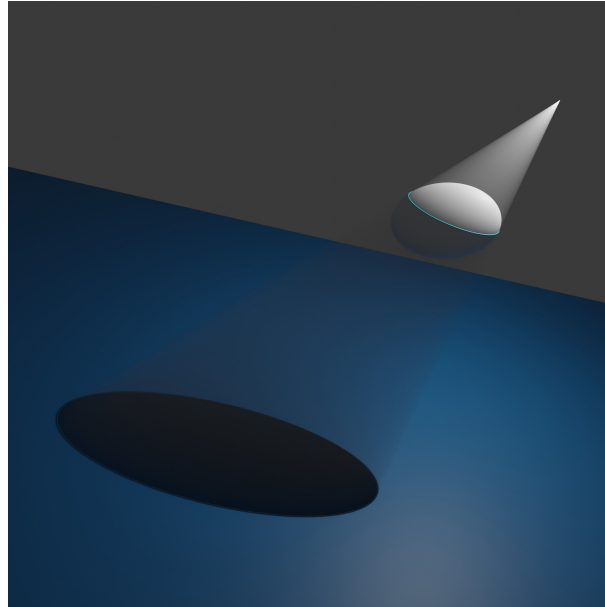


Figure 13. Shadow of an ellipsoid.

3.7 Polarity and tangent planes

Let $\mathcal{Q} \subset \mathbb{RP}^n$ be a quadric of signature (r, s, t) .

For a point $X = [x]$, its *polar hyperplane* (with respect to \mathcal{Q}) is given by

$$X^\perp = \{[y] \in \mathbb{RP}^n \mid b(x, y) = 0\}.$$

If the point X has signature

- ▶ $(+)$, then X^\perp has signature $(r - 1, s, t)$.
- ▶ $(-)$, then X^\perp has signature $(r, s - 1, t)$.
- ▶ (0) , then X^\perp has signature $(r - 1, s - 1, t + 1)$.

For the cases (+) and (-), we have established, that the intersection of X^\perp with \mathcal{Q} consists of all points common with the cone of contact \mathcal{C}_X .

In the case (0), every point $Y \in X^\perp$ that does not lie on the quadric is a tangent line of \mathcal{Q} . Thus for a point $X \in \mathcal{Q}$ on the quadric, the polar hyperplane is the plane containing (and spanned by) all tangent lines through X , which we call the *tangent plane* of \mathcal{Q} in the point X .

Example 3.7 (Tangent planes of a hyperboloid). Consider a one-sheeted hyperboloid $\mathcal{H} \subset \mathbb{RP}^3$, i.e. a quadric of signature $(+ + - -)$. Then a tangent plane X^\perp in any point $X \in \mathcal{H}$ has signature $(+ - 0)$. Thus, the restriction of \mathcal{H} to X^\perp consists of two lines.

In particular this means, that a one-sheeted hyperboloid, contains two lines through every point. In fact, it is a doubly ruled surface, and contains two families of lines, called its *generators*.

Example 3.8 (Projection of a generator).

What is the shadow of a generator of a hyperboloid?

Consider a one-sheeted hyperboloid $\mathcal{H} \subset \mathbb{RP}^3$, a generator $\ell \subset \mathcal{H}$, and a center of projection X not on \mathcal{H} . We consider the projection to X^\perp .

The projection of \mathcal{H} to X^\perp is given by a conic section

$$\mathcal{D} := \mathcal{C}_X \cap X^\perp = \mathcal{H} \cap X^\perp$$

of signature $(+ + -)$. Its affine type can be determined in a similar way to Example 3.6.

Denote the central projection of ℓ to X^\perp by $\tilde{\ell}$. The line ℓ intersects X^\perp in some point $A \in \mathcal{D}$, which is fixed under the projection to X^\perp . Thus, $A \in \tilde{\ell}$.

Assume there exists another point $B \in \ell$ such that its projection \tilde{B} lies on \mathcal{D} . Then the line $X \vee \tilde{B}$ is a tangent line of \mathcal{H} . On the other hand, this line intersects \mathcal{H} in the two distinct points B and \tilde{B} , which is a contradiction. Thus, the projection ℓ only intersects \mathcal{D} in A , and therefore is a tangent line of \mathcal{D} .

Note that projection to any other plane preserves this property.

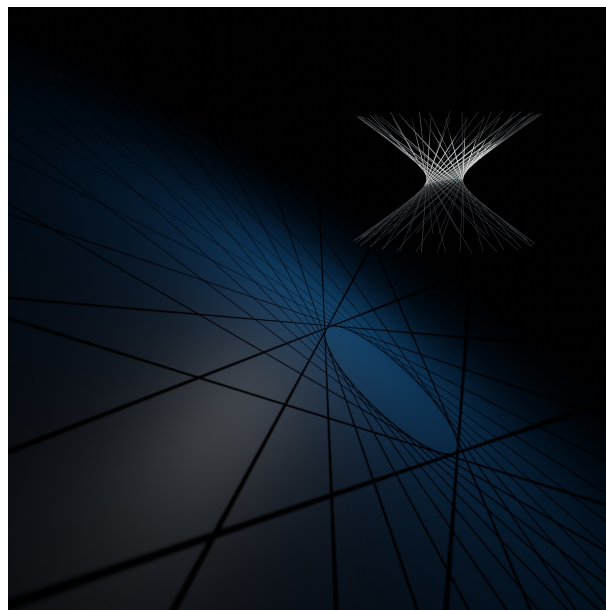


Figure 14. Shadow of the generators of a hyperboloid.

Differential geometric tangent plane Let us compare the notion of *tangent plane* that we have introduced for quadrics to the corresponding notion from Differential Geometry. In affine coordinates, we can view a quadric as a submanifold of \mathbb{R}^n given as a level set of the function

$$0 = x^T Q x = \begin{pmatrix} u^T & 1 \end{pmatrix} \left(\frac{S}{q^T} \middle| \frac{q}{\sigma} \right) \begin{pmatrix} u \\ 1 \end{pmatrix} = u^T S u + 2q^T u + \sigma =: f(u)$$

Then the normal vector of the tangent plane at some point $u_0 \in \mathbb{R}^n$ with $f(u_0) = 0$ is given by the gradient

$$\nabla_u f(u_0) = 2S u_0 + 2q.$$

Thus, the tangent plane at $u_0 \in \mathbb{R}$ is given by

$$\{u \in \mathbb{R}^n \mid \langle S u_0 + q, u - u_0 \rangle = 0\}$$

With

$$\langle S u_0 + q, u - u_0 \rangle = u_0^T S u + q^T u - u_0^T S u_0 - q^T u_0 = u_0^T S u + q^T u + q^T u_0 + \sigma$$

this coincides with the polar plane at u_0 in affine coordinates.

4 Pencils of quadrics

Definition 4.1. A projective subspace in the space of quadrics $\mathbf{P} \operatorname{Sym}(V)$ is called a *linear system of quadrics*. A linear system of quadrics is called *degenerate* if it solely consists of degenerate quadrics.

All quadrics through k generic points in $\mathbf{P}(V)$ form a linear system of quadrics of codimension k .

Example 4.1. The space of conics in \mathbb{RP}^2 is a 5-dimensional projective space

$$\mathbf{P} \operatorname{Sym}(\mathbb{R}^3) \cong \mathbb{RP}^5.$$

In homogeneous coordinates $[x] = [x_1, x_2, x_3]$ on \mathbb{RP}^2 and the corresponding homogeneous coordinates $\mathcal{Q} = [q_{11}, q_{22}, q_{33}, q_{12}, q_{23}, q_{13}]$ on the space of conics the equation for the point $[x]$ lying on the conic \mathcal{Q} is given by

$$q_{11}x_1^2 + q_{22}x_2^2 + q_{33}x_3^2 + q_{12}x_1x_2 + q_{23}x_2x_3 + q_{13}x_1x_3 = 0.$$

Let $X_1, X_2, X_3, X_4 \in \mathbb{RP}^2$ be four points in general position. Consider the set \mathcal{P} of all conics containing these four points. To explicitly describe this family of conics we simply the corresponding equations by choosing homogeneous coordinates such that

$$X_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad X_2 = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \quad X_3 = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}, \quad X_4 = \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix}$$

Then the representative matrices $Q = (q_{ij})_{1 \leq i \leq j \leq 3}$ for the conics in \mathcal{P} must satisfy

$$\begin{aligned} q_{11} + q_{22} + q_{33} + 2q_{12} + 2q_{23} + 2q_{13} &= 0 \\ q_{11} + q_{22} + q_{33} - 2q_{12} + 2q_{23} - 2q_{13} &= 0 \\ q_{11} + q_{22} + q_{33} + 2q_{12} - 2q_{23} - 2q_{13} &= 0 \\ q_{11} + q_{22} + q_{33} - 2q_{12} - 2q_{23} + 2q_{13} &= 0 \end{aligned}$$

By subtracting equations we obtain

$$q_{12} + q_{13} = 0, \quad q_{13} - q_{23} = 0, \quad q_{12} - q_{13} = 0,$$

which implies $q_{12} = q_{13} = q_{23} = 0$. By adding up all four equations we additionally obtain

$$q_{11} + q_{22} + q_{33} = 0$$

Thus, every conic in \mathcal{P} is given by

$$Q = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & -\lambda - \mu \end{pmatrix}$$

for some $[\lambda, \mu] \in \mathbb{RP}^1$, which describes a one-dimensional projective subspace in \mathbb{RP}^5 and thus a pencil of conics. The equations of the conics in this pencil are given by

$$\lambda(x_1^2 - x_3^2) + \mu(x_2^2 - x_3^2) = 0.$$

A linear system of quadrics of dimension 1, such as the one considered in Example 4.1, is called a *pencil of quadrics*.

Definition 4.2 (pencil of quadrics). A one-parameter family of quadrics in \mathbb{RP}^n that corresponds to a line in the space of quadrics $\text{PSym}(V)$ is called a *pencil of quadrics*.

A pencil of quadrics is called *non-degenerate* if not all quadrics in the pencil are degenerate.

Any two quadrics $\mathcal{Q}_1, \mathcal{Q}_2 \in \text{PSym}(V)$ span a pencil, which is given in homogeneous coordinates by

$$\mathcal{Q}_1 \vee \mathcal{Q}_2 = [\lambda \mathcal{Q}_1 + \mu \mathcal{Q}_2]_{[\lambda, \mu] \in \mathbb{RP}^1}.$$

Lemma 4.1. *A point which is contained in two quadrics of a pencil is contained in every quadric of that pencil.*

Proof. Exercise. □

Definition 4.3 (base point). A point which is contained in two (and thus every) quadric of a pencil of quadric is called a *base point* of that pencil.

Example 4.2. The pencil of conics in Example 4.1 has four base points.

Degenerate quadrics of a pencil

The degenerate quadrics of the pencil are characterized by the equation

$$\det(\lambda \mathcal{Q}_1 + \mu \mathcal{Q}_2) = 0.$$

If the pencil is not degenerate, we may assume \mathcal{Q}_2 is non-singular and set $\lambda = 1$. This leads to

$$\det(\mathcal{Q}_1 + \mu \mathcal{Q}_2) = 0,$$

which now is a polynomial equation in μ of order at most $n + 1$. Note that over \mathbb{C} it has exactly $n + 1$ solutions counting multiplicities.

Proposition 4.2. *A non-degenerate pencil of quadrics contains at most $n + 1$ degenerate quadrics. Over \mathbb{C} the multiplicities of the degenerate quadrics add up to $n + 1$.*

Example 4.3. The degenerate conics in the pencil of conics from Example 4.1 are given by

$$x_1^2 - x_3^2 = 0, \quad x_2^2 - x_3^2 = 0, \quad x_1^2 - x_2^2 = 0,$$

which each consists of a pair of opposite lines from the complete quadrangle defined by the four base points. They all have rank 2 and multiplicity 1.

Note that the diagonal triangle

$$A = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

of the complete quadrangle of base points is a *polar triangle* for all conics of the pencil, i.e., each point is the pole of the opposite line.

Some geometric properties of pencils

Proposition 4.3. *Let \mathcal{P} be a pencil of quadrics. Let X be a point and H a hyperplane containing the point X . If two quadrics from \mathcal{P} are tangent to H in X , then X is a base point of \mathcal{P} and all quadrics from \mathcal{P} are tangent to H in X .*

Proof. Exercise. □

Example 4.4. Not every pencil of conics is given by all conics through four given points (such as Example 4.1) as this example shows.

Let $X_1, X_2 \in \mathbb{RP}^2$ be two (distinct) points and $\ell_1, \ell_2 \subset \mathbb{RP}^2$ two (distinct) lines such that X_1 lies on ℓ_1 and X_2 lies on ℓ_2 . Consider the set \mathcal{P} of conics which are tangent to ℓ_1 in X_1 and to ℓ_2 in X_2 . We will show that \mathcal{P} is a non-degenerate pencil with base points X_1, X_2 and two degenerate conics, one of rank 2 and multiplicity 2 and one of rank 1 and multiplicity 1.

Choose homogeneous coordinates such that

$$X_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad X_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \quad \ell_1 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}^*, \quad \ell_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}^*.$$

The two tangency conditions are given by

$$Q \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \sim \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \quad Q \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \sim \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix},$$

which yields

$$\begin{aligned} q_{12} + q_{23} &= 0 \\ q_{12} + q_{23} &= 0 \\ q_{11} + q_{33} + 2q_{13} &= 0 \\ q_{11} - q_{33} + 2q_{13} &= 0, \end{aligned}$$

or equivalently,

$$q_{12} = q_{23} = q_{13} = 0, \quad q_{11} = q_{33}.$$

Thus, all conics from \mathcal{P} are given by

$$Q = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & -\lambda \end{pmatrix}$$

for some $[\lambda, \mu] \in \mathbb{RP}^1$. The equations of the conics in this pencil are given by

$$\lambda(x_1^2 - x_3^2) + \mu x_2^2 = 0.$$

Its degenerate quadrics are given by

$$x_1^2 - x_3^2 = 0,$$

which has multiplicity 1 and consists of the two lines ℓ_1, ℓ_2 , and

$$x_2 = 0,$$

which has multiplicity 2 and consists of the (double) line $X_1 \vee X_2$.

Proposition 4.4. *Let \mathcal{P} be a pencil of quadrics. Let H a hyperplane tangent to two quadrics of \mathcal{P} in the two points X, Y . Then X and Y are conjugate with respect to all quadrics in the pencil.*

Proof. Exercise. □

4.1 Classification of pencils of conics

A classification of pencils of conics can be achieved by investigating base points (number and multiplicities).

Proposition 4.5. *A non-degenerate pencil of conics has at most four base points.*

Proof. Assume the pencil has five base points. As stated in Example 3.4, five points (no four of which are on a line) determine a unique conic. If three of them would lie on a line, by Lemma 3.1 and Lemma 4.1, the entire line would be contained in every conic of the pencil, which contradicts that the pencil is non-degenerate. □

Determining the base points of a pencil of conics

Let

$$\mathcal{P} = \mathcal{Q}_1 \vee \mathcal{Q}_2$$

be a non-degenerate pencil of conics. Since \mathcal{P} is non-degenerate we may assume \mathcal{Q}_1 is non-degenerate and choose homogeneous coordinates in which its equation is given by¹

$$x_1^2 - x_2x_3 = 0.$$

The conic \mathcal{Q}_2 is given by

$$q_{11}x_1^2 + q_{22}x_2^2 + q_{33}x_3^2 + q_{12}x_1x_2 + q_{23}x_2x_3 + q_{13}x_1x_3 = 0.$$

The point

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \in \mathcal{Q}_1$$

is the only point of \mathcal{Q}_1 on the line $x_3 = 0$. We can further assume that²

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \notin \mathcal{Q}_2,$$

¹This is possible since every pencil contains at least one conic of signature $(++-)$.

²This is possible since any three points on a conic can be mapped to any other three points while preserving the conic.

or equivalently,

$$q_{22} \neq 0.$$

Thus, we can introduce affine coordinates

$$x = \frac{x_1}{x_3}, \quad y = \frac{x_2}{x_3}$$

without having any base points on the line at infinity. In affine coordinates the two equations for $\mathcal{Q}_1, \mathcal{Q}_2$ are give by

$$\begin{aligned} y &= x^2 \\ q_{11}x^2 + q_{22}y^2 + 2q_{12}xy + 2q_{23}y + 2q_{13}x + q_{33} &= 0. \end{aligned} \tag{6}$$

Substituting the first equation into the second we obtain

$$q_{22}x^4 + 2q_{12}x^3 + (q_{11} + 2q_{23})x^2 + 2q_{13}x + q_{33} = 0. \tag{7}$$

and every solution of (7) corresponds to exactly one solution of (6). One can assign the multiplicities of the roots to the base points of the pencil.

Over \mathbb{C} equation (7) has exactly 4 solutions counting multiplicities. Thus, for a non-degenerate pencil in \mathbb{CP}^2 there are exactly five possible cases, which we denote as follows:

- (I) four simple base points $(1, 1, 1, 1)$
- (II) one double and two simple base points $(2, 1, 1)$
- (III) two double base points $(2, 2)$
- (IV) one triple and one simple base point $(3, 1)$
- (V) one quadruple base point (4)

Using this one can prove the following classification result for pencils of conics in \mathbb{CP}^2 .

Theorem 4.6. *Two non-degenerate pencils of conics in \mathbb{CP}^2 are projectively equivalent if and only if they are of the same type.*

Furthermore, the degenerate conics, their multiplicities, and a normal form for each type are as stated in Table 15.

Type	Base points	Deg. conics	Normal form
I	1, 1, 1, 1	\times, \times, \times	$\lambda(x_1^2 - x_3^2) + \mu(x_2^2 - x_3^2) = 0$
II	2, 1, 1	$2\times, \times$	$\lambda(x_1^2 - x_2^2) + \mu x_2(x_2 - x_3) = 0$
III	2, 2	$2\parallel, \times$	$\lambda(x_1^2 - x_3^2) + \mu x_2^2 = 0$
IV	3, 1	$3\times$	$\lambda(x_1^2 - x_2x_3) + \mu x_1x_2 = 0$
V	4	$3\parallel$	$\lambda(x_1^2 - x_2x_3) + \mu x_2^2 = 0$

Figure 15. The classification of pencils of conics in \mathbb{CP}^2 . The two types of degenerate conics are two lines (\times), and a double line (\parallel).

Type	Pencil	Dual Pencil
I		
II		
III		
IV		
V		

Figure 16. Primal pencils of types I-V and the corresponding dual pencils.

Remark 4.1. From Table 15 we see that the types of pencils can also be characterized by the number and rank of their degenerate conics.

By complexification of all conics in a pencil the number of base points (counting multiplicities) of a real pencil is still 4. Yet some base points may be imaginary, which

always come in complex conjugate pairs. Thus, some of the complex cases split into multiple real cases:

- (Ia) four simple real base points $(1, 1, 1, 1)$
- (Ib) two simple real base points and a pair of simple imaginary base points $(1, 1, (1, \bar{1}))$
- (Ic) two pairs of simple imaginary base points $((1, \bar{1}), (1, \bar{1}))$
- (IIa) one double and two simple real base points $(2, 1, 1)$
- (IIb) one double real base point and a pair of simple imaginary base points $(2, (1, \bar{1}))$
- (IIIa) two double real base points $(2, 2)$
- (IIIb) a pair of double imaginary base points $(2, \bar{2})$
- (IV) one triple and one simple real base point $(3, 1)$
- (V) one quadruple real base point (4)

Similarly, this leads to the following classification result for pencils of conics in \mathbb{RP}^2 .

Theorem 4.7. *Two non-degenerate pencils of conics in \mathbb{RP}^2 are projectively equivalent if and only if they are of the same (real) type.*

Furthermore, the degenerate conics, their multiplicities, and a normal form for each type are as stated in Table 17.

Type	base points	# real	Deg. conics	Roots	Normal forms
Ia	$1, 1, 1, 1$	4	\times, \times, \times	$1, 1, 1$	$\lambda(x_1^2 - x_3^2) + \mu(x_2^2 - x_3^2) = 0$
Ib	$1, 1, (1, \bar{1})$	2	$\times, \circ, \bar{\circ}$	$1, (1, \bar{1})$	$\lambda(x_1^2 + x_2^2 - x_3^2) + \mu x_2 x_3 = 0$
Ic	$(1, \bar{1}), (1, \bar{1})$	0	\times, \bullet, \bullet	$1, 1, 1$	$\lambda(x_1^2 + x_2^2 + x_3^2) + \mu x_1 x_3 = 0$
IIa	$2, 1, 1$	3	$2\times, \times$	$2, 1$	$\lambda(x_1^2 - x_2^2) + \mu x_2(x_2 - x_3) = 0$
IIb	$2, (1, \bar{1})$	1	$2\bullet, \times$	$2, 1$	$\lambda(x_1^2 + x_2^2) + \mu x_2 x_3 = 0$
IIIa	$2, 2$	2	$2\parallel, \times$	$2, 1$	$\lambda(x_1^2 - x_3^2) + \mu x_2^2 = 0$
IIIb	$(2, \bar{2})$	0	$2\parallel, \bullet$	$2, 1$	$\lambda(x_1^2 + x_2^2) + \mu x_3^2 = 0$
IV	$3, 1$	2	$3\times$	3	$\lambda(x_1^2 - x_2 x_3) + \mu x_1 x_2 = 0$
V	4	1	$3\parallel$	3	$\lambda(x_1^2 - x_2 x_3) + \mu x_2^2 = 0$

Figure 17. The classification of real pencils of conics. There exist four different types of degenerate conics. (\times) Two real intersecting lines. (\circ) Two non-intersecting complex lines. (\bullet) Two complex conjugate lines which intersect in a real point. (\parallel) A real double line.

4.2 Classification of pencils of quadrics

The classification of pencils of conics by number and multiplicity of base points as discussed in Section 4.1 is specific to the 2-dimensional case. In higher dimensions the base points in general do not consist of a finite amount of points anymore, but constitute a subvariety of codimension 2.

Definition 4.4. Let $\mathcal{P} \subset \text{PSym}(\mathbb{C}^{n+1})$ be a pencil of quadrics. Let Q_1, Q_0 be two quadrics in \mathcal{P} with representative matrices $Q_1, Q_0 \in \text{Sym}(\mathbb{C}^{n+1})$. Then we call $Q_1\lambda + Q_0 \in \mathbb{C}[\lambda]^{(n+1) \times (n+1)}$ a *characteristic matrix* of \mathcal{P} .

A characteristic matrix uniquely determines its pencil together with the two quadrics spanning it. Vice versa, two characteristic matrices

$$Q_1\lambda + Q_0, \quad \text{and} \quad \tilde{Q}_1\lambda + \tilde{Q}_0,$$

describe the same pencil if and only if

$$\begin{aligned} \tilde{Q}_1 &= aQ_1 + cQ_0, \\ \tilde{Q}_0 &= bQ_1 + dQ_0, \end{aligned}$$

with $a, b, c, d \in \mathbb{C}$, $ad - bc \neq 0$. And thus, the corresponding values of λ are related by

$$\lambda = \frac{a\tilde{\lambda} + b}{c\tilde{\lambda} + d},$$

i.e. by a 1-dimensional projective transformation.

Now consider a projective transformation $f = [F] : \mathbb{CP}^n \rightarrow \mathbb{CP}^n$. It maps the pencil \mathcal{P} to $f(\mathcal{P})$ by acting on the characteristic matrix $Q_1\lambda + Q_0$ as

$$F^\top(Q_1\lambda + Q_0)F = F^\top Q_1 F \lambda + F^\top Q_0 F.$$

Thus, two pencils \mathcal{P} and $\tilde{\mathcal{P}}$ are *projectively equivalent*, i.e. related by a projective transformation, if and only if there exist characteristic matrices $\lambda Q_1 + Q_0$ and $\lambda \tilde{Q}_1 + \tilde{Q}_0$ such that Q_1 and Q_0 are simultaneously congruent to \tilde{Q}_1 and \tilde{Q}_0 , i.e.

$$\tilde{Q}_1 = F^\top Q_1 F, \quad \tilde{Q}_0 = F^\top Q_0 F$$

for some $F \in \text{GL}(n+1, \mathbb{C})$. Note that this does not mean that for two projectively equivalent pencils any pair of characteristic matrices is related by a simultaneously congruence.

For a classification, we should find a sufficient number of invariants of pencils under projective transformations and under a change of characteristic matrices. Firstly note that rank of a quadric is invariant under projective transformations, and thus degenerate quadrics are mapped to degenerate quadrics. Furthermore, the multiplicities μ_1, \dots, μ_s of the degenerate quadrics given by

$$\det(\lambda Q_1 + Q_0) = c(\lambda - \lambda_1)^{\mu_1} \cdots (\lambda - \lambda_s)^{\mu_s}, \quad \sum_{i=1}^s \mu_i = n+1$$

In the 2-dimensional case (classification of pencils of conics, see Table 15 and 17) it turns out that rank and multiplicities of the degenerate conics is indeed sufficient to characterize each equivalence class, and thus lead to a full classification. However, in higher dimensions this information is still insufficient.

A closer investigation of how the rank drops for each degenerate quadric leads to additional invariants.

Example 4.5.

(i) Pencil of conics with 3 degenerate conics:

$$\lambda Q_1 + Q_0 = \begin{pmatrix} \lambda - \lambda_1 & 0 & 0 \\ 0 & \lambda - \lambda_2 & 0 \\ 0 & 0 & \lambda - \lambda_3 \end{pmatrix}$$

The degenerate conic and their multiplicities are given by

$$\det(\lambda Q_1 + Q_0) = (\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3) = 0$$

They have at most rank 2. Its non-trivial 2×2 -minors are given by

$$(\lambda - \lambda_1)(\lambda - \lambda_2), (\lambda - \lambda_2)(\lambda - \lambda_3), (\lambda - \lambda_3)(\lambda - \lambda_1),$$

and its monic greatest common divisor by 1. Thus, in particular, the pencil contains no conic of rank 1.

(ii) Pencil of conics with 2 degenerate conics:

$$\lambda Q_1 + Q_0 = \begin{pmatrix} \lambda - \lambda_1 & 0 & 0 \\ 0 & \lambda - \lambda_1 & 0 \\ 0 & 0 & \lambda - \lambda_2 \end{pmatrix}$$

The degenerate conic and their multiplicities are given by

$$\det(\lambda Q_1 + Q_0) = (\lambda - \lambda_1)^2(\lambda - \lambda_2) = 0$$

They have at most rank 2. Its non-trivial 2×2 -minors are given by

$$(\lambda - \lambda_1)^2, (\lambda - \lambda_1)(\lambda - \lambda_2)$$

and its monic greatest common divisor by $\lambda - \lambda_1$. Thus, in particular, for $\lambda = \lambda_1$ the rank drops down to 1.

(iii) Another pencil of conics with 2 degenerate conics:

$$\lambda Q_1 + Q_0 = \begin{pmatrix} 0 & \lambda - \lambda_1 & 0 \\ \lambda - \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda - \lambda_2 \end{pmatrix}$$

The degenerate conic and their multiplicities are given by

$$\det(\lambda Q_1 + Q_0) = -(\lambda - \lambda_1)^2(\lambda - \lambda_2)$$

They have at most rank 2. Its non-trivial 2×2 -minors are given by

$$-(\lambda - \lambda_1)^2, (\lambda - \lambda_1)(\lambda - \lambda_2), \lambda - \lambda_2$$

and its monic greatest common divisor by 1. Thus, in particular, the pencil contains no conics of rank 1.

Definition 4.5. Let $A \in \mathbb{C}[\lambda]^{(n+1) \times (n+1)}$ be a square polynomial matrix of rank $\ell = \text{rk } A$. Then for $k = 1, \dots, \ell$ the monic³ greatest common divisor D_k of all $k \times k$ minors of A is called the k -th *minor divisor* of A . We also define $D_0 := 1$.

Lemma 4.8. D_k divides D_{k+1} for $k = 0, \dots, \ell - 1$.

³Monic polynomials are polynomials with leading coefficient equal to 1.

For the minor divisors of the characteristic matrix $\lambda Q_1 + Q_0$ of a pencil we obtain

$$\begin{aligned}
D_{n+1} &= (\lambda - \lambda_1)^{\mu_{1,n+1}} \cdots (\lambda - \lambda_s)^{\mu_{s,n+1}} = c \det(\lambda Q_1 + Q_0), \\
&\vdots \\
D_k &= (\lambda - \lambda_1)^{\mu_{1k}} \cdots (\lambda - \lambda_s)^{\mu_{sk}} \\
D_{k+1} &= (\lambda - \lambda_1)^{\mu_{1,k+1}} \cdots (\lambda - \lambda_s)^{\mu_{s,k+1}} \\
&\vdots \\
D_0 &= 1.
\end{aligned}$$

The collection of multiplicities μ_{ij} are invariants for the pencil. By Lemma 4.8, the sequences $\mu_{i,n+1}, \dots, \mu_{i,0}$ are decreasing, and instead of the multiplicities μ_{ij} it is common to use their differences

$$\nu_{ij} := \mu_{ij} - \mu_{i,j-1},$$

which satisfy

$$\sum_{j=1}^{n+1} \nu_{ij} = \mu_{i,n+1}.$$

Together they constitute the *Segre symbol* of $\lambda Q_1 + Q_0$

$$[(\lambda_1 : \nu_{1,1}, \dots, \nu_{1,n+1}), \dots, (\lambda_s : \nu_{s,1}, \dots, \nu_{s,n+1})]$$

where ν_{ij} equal to zero are omitted, and often so are the roots λ_i .

Example 4.6. The Segre symbols for Example 4.5 are given by:

- (i) $[(\lambda_1 : 1), (\lambda_1 : 1), (\lambda_3 : 1)]$ which is abbreviated to $[111]$.
- (ii) $[(\lambda_1 : 1, 1), (\lambda_1 : 1)]$ which is abbreviated to $[(11)1]$.
- (iii) $[(\lambda_1 : 1, 2), (\lambda_1 : 1)]$ which is abbreviated to $[21]$.
- (iv) The only other two possible Segre symbols in the 2-dimensional case are given by $[(3)]$ and $[(21)]$, which leads to the 5 classes of pencils of conics in \mathbb{CP}^2 as seen in Section 4.1.

The Segre symbol may be used to obtain a full classification of pencils of quadrics in \mathbb{CP}^n .

Theorem 4.9. *Two non-degenerate pencils of quadrics in \mathbb{CP}^n are projectively equivalent if and only if they have the same Segre symbol up to a (complex) projective transformation of the roots.*

In practice the Segre symbol of a given pencil may be obtained by the following normal form.

Theorem 4.10. *Let $Q_1, Q_0 \in \text{Sym}(\mathbb{C}^{n+1})$ with Q_1 non-singular, and let*

$$J = \text{diag}(J_1, \dots, J_m), \quad J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & 1 & \\ & & & \lambda_i \end{pmatrix}$$

be the (complex) Jordan normal form of $Q_1^{-1}Q_0$. Then Q_1, Q_0 are simultaneously congruent via a (complex) congruence transformation to

$$\begin{aligned}\tilde{Q}_1 &= \text{diag}(E_1, \dots, E_m), \\ \tilde{Q}_0 &= \text{diag}(E_1 J_1, \dots, E_m J_m),\end{aligned}$$

where

$$E_i = \begin{pmatrix} & & 1 \\ & \ddots & \\ 1 & & \end{pmatrix}, \quad E_i J_i = \begin{pmatrix} & & \lambda_i \\ & \ddots & 1 \\ \lambda_i & & 1 \end{pmatrix}, \quad \dim E_i = \dim J_i.$$

It turns out that the sizes of the Jordan blocks are exactly the ν_{ij} if the corresponding Segre symbol. Note how the three pencils given in Example 4.5 are already in normal form, and the Segre symbols can be read off immediately.

In a similar way, the following real version of Theorem 4.10 can be used for a classification of pencils in \mathbb{RP}^n .

Theorem 4.11. *Let $Q_1, Q_0 \in \text{Sym}(\mathbb{R}^{n+1})$ with Q_1 non-singular, and let*

$$J = \text{diag}(J_1, \dots, J_r, J_{r+1}, \dots, J_m)$$

be the (real) Jordan normal form of $Q_1^{-1}Q_0$, where

$$J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & 1 & \\ & & & \lambda_i \end{pmatrix}, \quad i = 1, \dots, r$$

are the Jordan blocks for real eigenvalues $\lambda_1, \dots, \lambda_r$, and

$$J_j = \begin{pmatrix} \Lambda_j & I_2 & & \\ & \ddots & \ddots & \\ & & I_2 & \\ & & & \Lambda_j \end{pmatrix}, \quad \Lambda_j = \begin{pmatrix} a_j & -b_j \\ b_j & a_j \end{pmatrix}, \quad j = r+1, \dots, m$$

are the Jordan blocks for complex pairs of eigenvalues $\lambda_j = a_j + ib_j, \bar{\lambda}_j = a_j - ib_j, j = r+1, \dots, m$. Then Q_1, Q_0 are simultaneously congruent via a (real) congruence transformation to

$$\begin{aligned}\tilde{Q}_1 &= \text{diag}(\varepsilon_1 E_1, \dots, \varepsilon_r E_r, E_{r+1}, \dots, E_m), \\ \tilde{Q}_0 &= \text{diag}(\varepsilon_1 E_1 J_1, \dots, \varepsilon_r E_r J_r, E_{r+1} J_{r+1}, \dots, E_m J_m),\end{aligned}$$

where $\varepsilon = \pm 1$ (unique), and

$$E_i = \begin{pmatrix} & & 1 \\ & \ddots & \\ 1 & & \end{pmatrix}, \quad \dim E_i = \dim J_i.$$

The number of real eigenvalues together with the different possible signs ε_i account for different real subclasses of each complex class with a given Segre symbol. In the real classification one may also use invariants that encode the signature of the quadrics in the pencil, such as the index sequence and the signature sequence.

5 Fractals

Following [Fractal Geometry - Kenneth Falconer].

Example 5.1 (Cantor set). Start with the unit interval $[0, 1] \subset \mathbb{R}$ and consider a sequence of intervals, where in each step the middle third of all previous intervals is deleted.

$$\begin{aligned} F_0 &= [0, 1] \\ F_1 &= [0, \frac{1}{3}] \cup [\frac{2}{3}, 1] \\ F_2 &= [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1] \\ &\vdots \end{aligned}$$

This describes a decreasing sequence of sets, whose limit is called the *Cantor set*:

$$F = \bigcap_{k=0}^{\infty} F_k$$

It is an example of a compact uncountably infinite set without isolated points, that is nowhere dense in $[0, 1]$.

In the ternary (base 3) expansion of real numbers, deleting the middle third of each interval corresponds to deleting the numbers containing the digit 1. Thus, an alternative representation of the Cantor set is given by

$$F = \left\{ \sum_{i=1}^{\infty} a_i 3^{-i} \mid a_i \in \{0, 2\} \right\}.$$

Note that $\frac{1}{3}$ has the ternary expansion $0.1 = 0.0\bar{2}$ and therefore is captured by this description. Furthermore, note that the limit set F does not solely consist of boundary points of intervals in the sequence F_k , e.g. $\frac{1}{4} = 0.\bar{02} \in F$.

Defining the following two similarity transformations

$$S_1, S_2 : [0, 1] \rightarrow [0, 1], \quad S_1(x) = \frac{1}{3}x, \quad S_2(x) = \frac{1}{3}x + \frac{2}{3}$$

each step may alternatively be written as

$$F_k = S_1(F_{k-1}) \cup S_2(F_{k-1}).$$

This describes a self-similarity which is still present in the limit:

$$F = S_1(F) \cup S_2(F).$$

The total length of the intervals deleted is given by

$$\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \cdots = \sum_{k=0}^{\infty} \frac{2^k}{3^{k+1}} = \frac{1}{3} \frac{1}{1 - \frac{2}{3}} = 1.$$

Thus the length of F is 0. While the interior of F is empty, every point of F is a limit point.

The Cantor set exhibits the following properties, which are typical for “fractals”.

- (i) Defined in very simple ways, perhaps recursively.
- (ii) Has a fine structure, with detail on arbitrary scale.
- (iii) Too irregular to describe in traditional geometric terms (locally and globally).
Neither the locus of points that satisfy some simple geometric condition, nor the set of solutions of any simple equation.
- (iv) Some sort of self-similarity (possibly approximate or statistical).
- (v) The size is not quantified by usual measures such as length.
The “fractal dimension” (defined in some way) is greater than the topological dimension.

Example 5.2. Some other simple examples:

- (i) von Koch curve
- (ii) Cantor dust
- (iii) Sierpinski triangle

5.1 Iterated function systems

Definition 5.1. Let $D \subset \mathbb{R}^n$ be closed. A map $S : D \rightarrow D$ is called a *contraction* if there exists a $0 < r < 1$ such that for all $x, y \in D$

$$|S(x) - S(y)| \leq r |x - y|.$$

Remark 5.1.

- (i) Contractions are continuous.
- (ii) If $|S(x) - S(y)| \leq r |x - y|$, then S is a similarity transformations, which is called a *contracting similarity*.

Definition 5.2. An *iterated function system (IFS)* is a finite family of contractions $\{S_1, \dots, S_m\}$, $m \geq 2$.

A non-empty compact $F \subset D$ is called an *attractor* (or invariant set) of the IFS if

$$F = \bigcup_{i=1}^m S_i(F).$$

Remark 5.2. If the iterated function systems consists only of contracting similarities, then the attractor is called a *self-similar set*.

Example 5.3 (Cantor set). Let $D = [0, 1]$ and consider the two maps $S_1, S_2 : D \rightarrow D$,

$$S_1(x) = \frac{1}{3}x, \quad S_2(x) = \frac{1}{3}x + \frac{2}{3}.$$

Both are contracting similarities with $r = \frac{1}{3}$. Thus, $\{S_1, S_2\}$ is an IFS.

The Cantor set satisfies

$$F = S_1(F) \cup S_2(F),$$

and thus is an attractor of the IFS.

The fundamental property of IFS is that they determine a unique attractor, which is usually a fractal.

To this end, denote

$$\mathcal{C} = \{A \subset D \mid \emptyset \neq A \text{ compact}\},$$

and define the map

$$S : \mathcal{C} \rightarrow \mathcal{C}, \quad S(A) = \bigcup_{i=1}^m S_i(A).$$

Then $F \in \mathcal{C}$ is an attractor if and only if it is a fixed point of S , i.e.

$$S(F) = F.$$

Recall the Banach fixed point theorem:

Theorem 5.1 (Banach fixed point theorem). *Let (X, d) be a non-empty complete metric space and $T : X \rightarrow X$ a contraction. Then T has a unique fixed point $x^* \in X$.*

Moreover, for any $x_0 \in X$ the fixed point is given by

$$x^* = \lim_{k \rightarrow \infty} T^k(x_0).$$

To use this theorem, we equip \mathcal{C} with a metric.

Definition 5.3. For $A, B \in \mathcal{C}$ the *Hausdorff metric* is given by

$$d(A, B) := \inf \{\delta > 0 \mid A \subset B_\delta \text{ and } B \subset A_\delta\}$$

where

$$A_\delta := \{x \in D \mid |x - a| \leq \delta \text{ for some } a \in A\} = \bigcup_{a \in A} \overline{B_\delta(a)}.$$

is the (closed) δ -neighborhood of the set A .

Remark 5.3.

- (i) The quantity d is well-defined since A, B are bounded and thus the infimum exists (is finite).
- (ii) The Hausdorff metric may also be given in the following way

$$d(A, B) = \max \left\{ \max_{a \in A} d(a, B), \max_{b \in B} d(b, A) \right\}$$

$$\text{where } d(a, B) = \min_{b \in B} |a - b|.$$

Lemma 5.2. *The Hausdorff metric d is a complete metric on \mathcal{C} .*

Proof. Exercise.

d is a metric. Show:

- (i) $d(A, B) \geq 0$ and $d(A, B) = 0 \Leftrightarrow A = B$.
- (ii) $d(A, B) = d(B, A)$.
- (iii) $d(A, B) \leq d(A, C) + d(C, B)$.

d is complete. Show:

(iv) Every Cauchy sequence converges (in \mathcal{C}).

□

Lemma 5.3. *The map S is a contraction on (\mathcal{C}, d) .*

Proof. Let $0 < r_i < 1$, $i = 1, \dots, m$, such that

$$|S_i(x) - S_i(y)| \leq r_i |x - y|.$$

For $A, B \in \mathcal{C}$

$$S_i(A) \subset (S_i(B))_\delta \Rightarrow \bigcup_{i=1}^m S_i(A) \subset \left(\bigcup_{i=1}^m S_i(B) \right)_\delta,$$

and thus

$$d(S(A), S(B)) = d\left(\bigcup_{i=1}^m S_i(A), \bigcup_{i=1}^m S_i(B)\right) \leq \max_{i=1, \dots, m} d(S_i(A), S_i(B)) \leq \left(\max_{i=1, \dots, m} r_i\right) d(A, B).$$

□

Thus, by the Banach fixed point theorem, S has a unique fixed point $F \in \mathcal{C}$, and we obtain the following theorem on IFS:

Theorem 5.4. *Let $\{S_1, \dots, S_m\}$ be an IFS on $D \subset \mathbb{R}^n$. Then it has a unique attractor F , i.e. an $F \in \mathcal{C}$ such that*

$$S(F) = F.$$

Moreover, for any $E \in \mathcal{C}$

$$F = \lim_{k \rightarrow \infty} S^k(E).$$

We make the following observations:

- In every step the sequence $S^k(E)$ provides a better approximation of the attractor:

$$d(S^k(E), F) = d(S^k(E), S(F)) \leq c d(S^{k-1}(E), F) \leq \dots \leq c^k d(E, F),$$

where $c = \max_{i=1, \dots, m} r_i$.

- The approximation in the k -th step is given by

$$S^k(E) = \bigcup_{(i_1, \dots, i_k) \in \{1, \dots, m\}^k} S_{i_1} \circ \dots \circ S_{i_k}(E),$$

which is the union of m^k sets.

- To visualize an approximation of the attractor, each of the m^k sets of $S^k(E)$ may be drawn fully, or a representative points $S^k(x_0)$ with $x_0 \in E$ may be visualized.
- To obtain a statistical approximation of points in the attractor, we may draw the sequences (i_1, \dots, i_k) randomly. Then the sequence of points $x_k = S_{i_1} \circ \dots \circ S_{i_k}(x_0)$ may be drawn from a certain term onwards.

- If we chose an $E \in \mathcal{C}$ with $S(E) \subset E$, the sequence $S^k(E)$ is a decreasing sequence of sets with limit and thus

$$F = \lim_{k \rightarrow \infty} S^k(E) = \bigcap_{k=0}^{\infty} S^k(E).$$

Then for any $x \in F$ there exists a (not necessarily unique) sequence (i_1, \dots, i_k) such that $x \in S_{i_1} \circ \dots \circ S_{i_k}(E)$. This sequence provides a natural encoding for x by

$$x = \bigcap_{k=0}^{\infty} S_{i_1} \circ \dots \circ S_{i_k}(E).$$

5.2 Fractal dimensions

For $d \in \mathbb{N}$ consider a compact smooth d -dimensional submanifold \mathcal{M} of \mathbb{R}^n (a curve for $d = 1$, a surface for $d = 2$, ...) For $\delta > 0$, let N_δ be the (smallest) number of δ -boxes (cubes of side length δ) it takes to fully cover \mathcal{M} .

If we halve the side lengths of the cubes (considering $\frac{\delta}{2}$ -boxes), we expect the number of boxes it takes to cover \mathcal{M} to increase approximately by a factor 2^d . More generally, we expect the number N_δ to behave like

$$N_\delta \sim \frac{c}{\delta^d} \quad (8)$$

in the limit $\delta \rightarrow 0$, where c is some constant.

The dimension d can be recovered from (8) by taking logarithms

$$\log N_\delta \sim \log c - d \log \delta,$$

and the limit $\delta \rightarrow 0$

$$d \sim -\frac{\log N_\delta}{\log \delta} + \frac{\log c}{\log \delta} \rightarrow -\frac{\log N_\delta}{\log \delta}.$$

The behavior of N_δ given by (8) may equivalently be described by (without the need of the constant c)

$$N_\delta \delta^s \rightarrow \begin{cases} \infty & \text{if } s < d \\ 0 & \text{if } s > d \end{cases} \quad (\delta \rightarrow 0).$$

Thus, the function

$$f(s) = \lim_{\delta \rightarrow 0} N_\delta \delta^s$$

jumps from ∞ to 0 at the value $s = d$ of the dimension of \mathcal{M} .

We may use these ideas to define fractal dimensions for general (bounded) sets $F \subset \mathbb{R}^n$. To this end, let us generalize first from covers by δ -boxes to general δ -covers.

Definition 5.4. For $U \subset \mathbb{R}^n$ the *diameter* of U is given by

$$\text{diam}(U) = \sup_{x, y \in U} |x - y|$$

For $F \subset \mathbb{R}^n$ a countable (or finite) collection $(U_i)_{i=1}^\infty$ of sets of diameter at most $\delta > 0$ that cover F is called a δ -cover of F , i.e.,

$$F \subset \bigcup_{i=1}^{\infty} U_i, \quad \text{and} \quad \text{diam}(U_i) \leq \delta, \quad i = 1, \dots, \infty.$$

Before we get deeper into the different definitions of fractal dimensions we give a general overview of the ideas.

Box-counting dimension Let $F \subset \mathbb{R}^n$ be bounded. If we are only interested in the smallest number of sets in a δ -cover, it is sufficient to consider finite covers. Thus, let

$$N_\delta = \min \{N \mid (U_i)_{i=1}^N \text{ is a } \delta\text{-cover of } F\},$$

and if the limit exists define

$$\dim_B F = -\lim_{\delta \rightarrow 0} \frac{\log N_\delta}{\log \delta},$$

which is called the *box-counting dimension* of F . It is given by the value $s \geq 0$ where the function

$$f(s) = \lim_{\delta \rightarrow 0} N_\delta \delta^s = \begin{cases} \infty & \text{if } s < \dim_B F, \\ 0 & \text{if } s > \dim_B F. \end{cases}$$

jumps from ∞ to 0.

In general the limit $-\frac{\log N_\delta}{\log \delta}$ does not exist (only limit inferior and limit superior do), and there is a gap between the value ∞ and 0 of the function $f(s)$. However, the box-counting dimension leads to a definition of fractal dimension which is easy to approximate computationally.

Hausdorff dimension A mathematically more satisfying definition may be obtained by the following very similar idea. For $F \subset \mathbb{R}^n$ instead of $N_\delta \delta^s$, consider

$$\mathcal{H}_\delta^s = \inf \left\{ \sum_{i=1}^\infty \text{diam}(U_i)^s \mid (U_i)_{i=1}^\infty \text{ is a } \delta\text{-cover of } F \right\} \leq N_\delta \delta^s.$$

For $s \geq 0$ and in the limit $\delta \rightarrow 0$, the function

$$\mathcal{H}^s = \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^s = \begin{cases} \infty & \text{if } s < \dim_H F, \\ 0 & \text{if } s > \dim_H F. \end{cases}$$

jumps from ∞ to 0 at a well-defined value $\dim_H F$, which is called the *Hausdorff dimension* of F .

Note that from $\mathcal{H}_\delta^s \leq N_\delta \delta^s$ it follows that

$$\dim_H F \leq \dim_H B.$$

Moreover, \mathcal{H}^s defines a measure (on the Borel sets of \mathbb{R}^n), called the *Hausdorff measure*, which generalizes the Lebesgue measure. In particular, it satisfies the scaling property

$$\mathcal{H}^s(g(F)) = \lambda^s \mathcal{H}^s(F)$$

for any similarity transformation $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ with scaling factor $\lambda > 0$. However, the Hausdorff dimension as a fractal dimension is harder to estimate by computational methods than the box-counting dimension.

Similarity dimension For self-similar sets it is particularly easy to define a corresponding dimension. Thus, let $\{S_1, \dots, S_m\}$ be an iterated function system of similarity transformations

$$|S_i(x) - S_i(y)| = r_i |x - y|$$

with some $0 < r_i < 1$. Then its attractor

$$F = \bigcup_{i=1}^{\infty} S_i(F) \quad (9)$$

is a self-similar set. Assume this union in (9) is disjoint, and that $s \geq 0$ such that F has positive and finite Hausdorff measure $0 < \mathcal{H}^s < \infty$ (or any reasonable measure satisfying the scaling property).

$$\mathcal{H}^s(F) = \sum_{i=1}^m \mathcal{H}^s(S_i(F)) = \sum_{i=1}^m r_i^s \mathcal{H}^s(F),$$

which implies

$$\sum_{i=1}^m r_i^s = 1. \quad (10)$$

Thus, we may define

$$\dim_S F = s$$

to be the number s satisfying (10) for any self-similar set, which is called the *similarity dimension* of F .

In the case where the sets in (9) do not overlap too much (or the union is even disjoint) the similarity dimension satisfies

$$\dim_S F = \dim_H F = \dim_B F.$$

Note that, if all similarity transformations have the same scaling factor $r = r_i$ we obtain $mr^s = 1$, or equivalently,

$$\dim_S F = -\frac{\log m}{\log r}.$$

Example 5.4 (Cantor set). We compute the three introduced dimensions for the Cantor set $F \subset [0, 1]$. Recall that F_k consists of 2^k intervals of length $\frac{1}{3^k}$ which have distance at least $\frac{1}{3^k}$, and

$$F = \bigcap_{k=0}^{\infty} F_k.$$

- (i) **Box-counting dimension.** It is sufficient to consider the decreasing sequence $\delta_k = \frac{1}{3^k}$ (cf. Remark 5.5).

► The 2^k intervals of F_k provide a δ_k -cover for F . Thus, $N_{\delta_k} \leq 2^k$, and therefore

$$\frac{\log N_{\delta_k}}{-\log \delta_k} \leq \frac{\log 2^k}{-\log \frac{1}{3^k}} = \frac{k \log 2}{k \log 3} = \frac{\log 2}{\log 3}$$

- On the other hand, an interval of length δ_k intersects at most one of the 2^{k-1} intervals of F_{k-1} , each of which contains points from F . Thus, $N_{\delta_k} \geq 2^{k-1}$, and therefore

$$\frac{\log N_{\delta_k}}{-\log \delta_k} \geq \frac{\log 2^{k-1}}{-\log \frac{1}{3^k}} = \frac{(k-1) \log 2}{k \log 3} \rightarrow \frac{\log 2}{\log 3} \quad (k \rightarrow \infty)$$

Thus the box-counting dimension is given by

$$\dim_B F = \lim_{\delta \rightarrow 0} \frac{\log N_\delta}{-\log \delta} = \lim_{k \rightarrow \infty} \frac{\log N_{\delta_k}}{-\log \delta_k} = \frac{\log 2}{\log 3}.$$

(ii) **Hausdorff dimension.** Let $s = \frac{\log 2}{\log 3}$. We show that

$$\frac{1}{2} \leq \mathcal{H}^s \leq 1.$$

- Consider the sequence $\delta_k = \frac{1}{3^k}$. Then the 2^k intervals of F_k provide a δ_k -cover of F . And thus,

$$\mathcal{H}_{\delta_k}^s \leq \frac{2^k}{3^{ks}} = 1$$

and therefore $\mathcal{H}^s \leq 1$.

- To prove $\mathcal{H}^s \geq \frac{1}{2}$, we show

$$\sum_{i=1}^N \text{diam}(U_i)^s \geq \frac{1}{2} = \frac{1}{3^s}$$

for finite cover $(U_i)_{i=1}^N$ consisting of closed intervals in $[0, 1]$ (by the compactness of F the same statement then holds for arbitrary covers).

For each U_i let k_i be such that

$$\frac{1}{3^{k_i+1}} \leq \text{diam}(U_i) < \frac{1}{3^{k_i}}.$$

Then U_i intersects at most one of the intervals in F_{k_i} , and for $j \geq k_i$ it intersects at most

$$2^{j-k_i} = \frac{2^j}{2^{k_i}} = \frac{2^j}{3^{sk_i}} \leq 2^j 3^s \text{diam}(U_i)^s$$

of the intervals of F_j . Choose $j = \max_{i=1, \dots, N} k_i$. The cover $(U_i)_{i=1}^N$ intersects all 2^j intervals of F_j , and thus

$$2^j \leq \sum_{i=1}^N 2^{j-k_i} \leq \sum_{i=1}^N 2^j 3^s \text{diam}(U_i)^s$$

Thus, the Hausdorff dimension is given by

$$\dim_H F = s = \frac{\log 2}{\log 3}.$$

- (iii) **Similarity dimension.** The Cantor set is a self-similar set. With the two similarities

$$S_1(x) = \frac{x}{3}, \quad S_2(x) = \frac{x}{3} + \frac{2}{3},$$

it is given by

$$F = S_1(F) \cup S_2(F),$$

where the union is disjoint. The scaling factors of S_1 and S_2 are equal and given by $r = \frac{1}{3}$. Thus, the similarity dimension s of F is given by

$$2 \frac{1}{3^s} = 1,$$

which is equivalent to

$$\dim_S F = s = \frac{\log 2}{\log 3}.$$

It turns out, that for the Cantor set all three dimensions are equal and given by

$$\dim_S F = \dim_B F = \dim_H F = \frac{\log 2}{\log 3} \approx 0,630929754.$$

Properties of fractal dimensions Before we come to the formal definitions of the introduced fractal dimensions, we list some properties that we will encounter, and might possibly be considered desired properties of dimensions.

- (i) **Monotonicity.** $E \subset F \Rightarrow \dim E \leq \dim F$.
- (ii) **Range of values.** $F \subset \mathbb{R}^n \Rightarrow 0 \leq \dim F \leq n$.
- (iii) **Finite stability.** $\dim(E \cup F) = \max\{\dim E, \dim F\}$.
- (iv) **Countable stability.** $\dim(\bigcup_{i=1}^{\infty} F_i) = \sup_i \dim F_i$.
- (v) **Finite sets.** $\dim F = 0$ if F is finite.
- (vi) **Countable sets.** $\dim F = 0$ if F is countable.
- (vii) **Open sets.** $\dim F = n$ if F is an open subset of \mathbb{R}^n .
- (viii) **Smooth manifold.** $\dim F = d$ if F is a smooth d -dimensional submanifold of \mathbb{R}^n .
Coincidence with topological dimension on topological manifolds is not desired.
- (ix) **Geometric invariance.** $\dim f(F) = \dim F$ if f is a Euclidean, similarity, or affine transformation of \mathbb{R}^n .
Note that the fractal dimensions are not of a topological nature, but of a geometric nature, in particular involving the Euclidean metric.
- (x) **Lipschitz invariance.** $\dim f(F) = \dim F$ if f is a bi-Lipschitz map.
This property is stronger than geometric invariance, but encountered for many fractal dimensions.

5.2.1 Box-counting dimension

Definition 5.5 (Box-counting dimension). Let $F \subset \mathbb{R}^n$ be bounded, and for $\delta > 0$ let N_δ be the least number of sets of any δ -cover of F , i.e.

$$N_\delta = \min \left\{ N \mid (U_i)_{i=1}^N \text{ is a } \delta\text{-cover of } F \right\}.$$

Then

$$\underline{\dim}_B F = \lim_{\delta \rightarrow 0} \frac{\log N_\delta}{-\log \delta}, \quad \overline{\dim}_B F = \lim_{\delta \rightarrow 0} \frac{\log N_\delta}{-\log \delta}$$

are called the *lower and upper box-counting dimension* of F . If $\underline{\dim}_B F = \overline{\dim}_B F$, then the common value

$$\dim_B F = - \lim_{\delta \rightarrow 0} \frac{\log N_\delta}{\log \delta}$$

is called the *box-counting dimension* of F .

Remark 5.4. In the limit $\delta \rightarrow 0$, the quantity $N_\delta \delta^s$

$$\lim_{\delta \rightarrow 0} N_\delta \delta^s = \begin{cases} \infty & \text{if } s < \underline{\dim}_B F, \\ 0 & \text{if } s > \overline{\dim}_B F. \end{cases}$$

has a gap between the determined values of ∞ and 0 . This gap vanishes in the case where $\underline{\dim}_B F = \overline{\dim}_B F = \dim_B F$.

Remark 5.5. In the definition of box-counting dimensions it is enough to consider limits of decreasing sequences δ_k that satisfy

$$\delta_{k+1} \leq c \delta_k$$

for some $0 < c < 1$ (typically $\delta_k = c^k$). Indeed, for $\delta > 0$ with $\delta_{k+1} \leq \delta < \delta_k$, we find

$$\frac{\log N_\delta}{-\log \delta} \leq \frac{\log N_{\delta_{k+1}}}{-\log \delta_k} = \frac{\log N_{\delta_{k+1}}}{-\log \delta_{k+1} + \log \frac{\delta_{k+1}}{\delta_k}} \leq \frac{\log N_{\delta_{k+1}}}{-\log \delta_{k+1} + \log c},$$

and thus

$$\lim_{\delta \rightarrow 0} \frac{\log N_\delta}{-\log \delta} \leq \lim_{k \rightarrow \infty} \frac{\log N_{\delta_k}}{-\log \delta_k},$$

while the opposite inequality is trivially true for any subsequence. For the lower limit this is shown in the same way.

There are various equivalent characterizations of the box-counting dimensions.

Theorem 5.5. *The definition of box-counting dimensions is equivalent upon replacing the number N_δ by any of the following:*

- (i) *the smallest number of closed balls of radius δ that cover F .*
- (ii) *the smallest number of cubes of side length δ that cover F .*
- (iii) *the number of cubes in a δ -grid that intersect F (see Remark 5.6).*
- (iv) *the largest number of disjoint balls of radius δ and centers in F .*

Remark 5.6. A δ -grid of \mathbb{R}^n is a family of cubes of the form

$$[i_1\delta, (i_1 + 1)\delta] \times \cdots \times [i_n\delta, (i_n + 1)\delta]$$

with $(i_1, \dots, i_n) \in \mathbb{Z}^n$. This approach to box-counting dimension is the most convenient for computational approximation. By Remark 5.5 it is sufficient to consider decreasing sequences of δ_k -grids such as

$$\delta_k = \frac{1}{2^k},$$

halving the grid size in every step.

Remark 5.7. The characterization in (iv) uses packings by δ -balls instead of coverings by δ -balls. Packing and coverings are sometimes considered as being “dual”.

Another way of obtaining the box-counting dimension is by observing the change of volume of a set when it gets extruded. Consider the δ -neighborhood of an d -dimensional smooth manifold \mathcal{M}

$$\mathcal{M}_\delta = \bigcup_{x \in \mathcal{M}} \overline{B_\delta(x)}.$$

We can measure its n -dimensional volume using the Lebesgue measure \mathcal{L}^n on \mathbb{R}^n . Then the volume will behave like

$$\mathcal{L}^n(\mathcal{M}_\delta) \sim c\delta^{n-d}$$

in the limit $\delta \rightarrow 0$ with some constant c , which is a measure for the d -dimensional volume (length, area, ...) of \mathcal{M} . This is due to the Minkowski–Steiner formula. Thus, the dimension of \mathcal{M} may be recovered by

$$d = n - \lim_{\delta \rightarrow 0} \frac{\mathcal{L}^n(\mathcal{M}_\delta)}{\log \delta}.$$

Similarly, an arbitrary set $F \subset \mathbb{R}^n$ may be regarded as s -dimensional if

$$\mathcal{L}^n(F_\delta) \sim c\delta^{n-s}.$$

The so defined value s turns out to coincide with the box-counting dimension.

Theorem 5.6. *Let $F \subset \mathbb{R}^n$ be bounded. Then*

$$\underline{\dim}_B F = n - \lim_{\delta \rightarrow 0} \frac{\mathcal{L}^n(F_\delta)}{\log \delta}, \quad \overline{\dim}_B F = n - \overline{\lim}_{\delta \rightarrow 0} \frac{\mathcal{L}^n(F_\delta)}{\log \delta},$$

where F_δ is the δ -neighborhood of F and \mathcal{L}^n is the Lebesgue measure.

The box-counting dimensions are invariant under bi-Lipschitz map.

Theorem 5.7. *Let $F \subset \mathbb{R}^n$ be bounded.*

(i) *If $f : F \rightarrow \mathbb{R}^m$ is a Lipschitz map, i.e., there exists a constant $c > 0$ such that*

$$|f(x) - f(y)| \leq c|x - y|$$

for all $x, y \in F$, then

$$\underline{\dim}_B f(F) \leq \underline{\dim}_B F, \quad \text{and} \quad \overline{\dim}_B f(F) \leq \overline{\dim}_B F.$$

(ii) If $f : F \rightarrow \mathbb{R}^m$ is a bi-Lipschitz map, i.e., there exist constants $0 < c_1 \leq c_2$ such that

$$c_1 |x - y| \leq |f(x) - f(y)| \leq c_2 |x - y|$$

for all $x, y \in F$, then

$$\underline{\dim}_B f(F) = \underline{\dim}_B F, \quad \text{and} \quad \overline{\dim}_B f(F) = \overline{\dim}_B F.$$

Proof.

(i) If (U_i) is a δ -cover of F , then so is $(U_i \cap F)$. Then $(f(U_i \cap F))$ is a $c\delta$ -cover of $f(F)$, and thus $N_{c\delta}(f(F)) \leq N_\delta(F)$ for all $\delta > 0$. So

$$\frac{\log N_{c\delta}(f(F))}{-\log(c\delta) + \log c} \leq \frac{\log N_\delta(F)}{-\log \delta}.$$

Taking limes superior and inferior as $\delta \rightarrow 0$ gives the result.

(ii) Bi-Lipschitz maps are bijective. Apply (i) to f and f^{-1} .

□

Remark 5.8. The first statement of Theorem 5.7 may for example be used to show that the box-counting dimension is reduced under projection, while the second implies that it is invariant under affine transformations.

The following theorem summarizes properties of the box-counting dimension.

Theorem 5.8. *The box-counting dimensions have the following properties:*

(i) **Monotonicity.** If $E \subset F$, then

$$\underline{\dim}_B E \leq \underline{\dim}_B F \quad \text{and} \quad \overline{\dim}_B E \leq \overline{\dim}_B F.$$

(ii) **Range of values.**

$$0 \leq \underline{\dim}_B F \leq \overline{\dim}_B F \leq n.$$

(iii) **Finite stability.** $\overline{\dim}_B$ is finitely stable, i.e.,

$$\overline{\dim}_B(E \cup F) = \max\{\overline{\dim}_B E, \overline{\dim}_B F\}.$$

(v) **Finite sets.** $\dim_B F = 0$ if F is a finite set.

(vii) **Open sets.** $\dim_B F = n$ if F is an open subset of \mathbb{R}^n .

(viii) **Smooth manifold.** $\dim_B F = d$ if F is a smooth d -dimensional submanifold of \mathbb{R}^n .

(ix) **Geometric invariance.** See Remark 5.8.

(x) **Lipschitz invariance.** See Theorem 5.7.

Remark 5.9. Note that the finite stability only holds for $\overline{\dim}_B$, not for $\underline{\dim}_B$.

However, the following lemma leads to some undesired properties.

Lemma 5.9. *Let $F \subset \mathbb{R}^n$ and \bar{F} denote the closure of F (the smallest closed subset of \mathbb{R}^n containing F). Then*

$$\underline{\dim}_B \bar{F} = \underline{\dim}_B F, \quad \text{and} \quad \overline{\dim}_B \bar{F} = \overline{\dim}_B F.$$

Proof. The smallest number of closed δ -balls that cover F equals the smallest number of closed δ -balls that cover \bar{F} . \square

This means that for any subset $F \subset \mathbb{R}^n$ which is dense in an open region of \mathbb{R}^n , we have

$$\dim_B F = n.$$

In particular, for the countable set of rational numbers this implies

$$\dim_B \mathbb{Q} = \dim_B \bar{\mathbb{Q}} = \dim_B \mathbb{R} = 1,$$

and thus

(iv) **Not zero on countable sets.** \dim_B does in general not vanish on countable sets.

Furthermore, this implies that

$$1 = \dim_B \mathbb{Q} \neq \sup_{a \in \mathbb{Q}} \dim_B \{a\} = 0,$$

and thus, the box-counting dimension does not satisfy countable stability

(vi) **No countable stability.**

$$\dim_B \left(\bigcup_{i=1}^{\infty} F_i \right) \text{ is in general not equal to } \sup_i \dim_B F_i.$$

Example 5.5. Another example of a very sparse compact set with non-vanishing box-counting dimension is given by

$$F = \left\{ 0, 1, \frac{1}{2}, \frac{1}{3}, \dots \right\}, \quad \dim_B F = \frac{1}{2}.$$

5.2.2 Hausdorff dimension

First the Hausdorff measure is defined.

Definition 5.6. Let $F \subset \mathbb{R}^n$, and $s \geq 0$. Then we define

$$\mathcal{H}_\delta^s(F) = \inf \left\{ \sum_{i=1}^{\infty} \text{diam}(U_i)^s \mid (U_i)_{i=1}^{\infty} \text{ is a } \delta\text{-cover of } F \right\}$$

for any $\delta > 0$, and the (s -dimensional) Hausdorff measure of F by

$$\mathcal{H}^s(F) = \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^s$$

Remark 5.10. If δ decreases the class of δ -covers of F decreases, and so does the infimum \mathcal{H}_δ^s . Thus \mathcal{H}_δ^s decreases as $\delta \rightarrow 0$, and therefore the limit always exists.

Theorem 5.10. *The Hausdorff measure \mathcal{H}^s is an outer measure on \mathbb{R}^n :*

- (i) If $A \subset \mathbb{R}^n$, then $0 \leq \mathcal{H}^s(A) \leq \infty$.
- (ii) $\mathcal{H}^s(\emptyset) = 0$.
- (iii) if $A, B \subset \mathbb{R}^n$ with $A \subset B$, then $\mathcal{H}^s(A) \leq \mathcal{H}^s(B)$.
- (iv) If $(A_i)_{i=1}^\infty$ countable (or finite) sequence of sets in \mathbb{R}^n , then

$$\mathcal{H}^s \left(\bigcup_{i=1}^\infty A_i \right) \leq \sum_{i=1}^\infty \mathcal{H}^s(A_i).$$

Furhermore, \mathcal{H}^s defines a measure on the Borel sets of \mathbb{R}^n (or more generally on \mathcal{H}^s -measurable sets):

- (v) If $(A_i)_{i=1}^\infty$ countable (or finite) sequence of disjoint Borel sets in \mathbb{R}^n , then

$$\mathcal{H}^s \left(\bigcup_{i=1}^\infty A_i \right) = \sum_{i=1}^\infty \mathcal{H}^s(A_i).$$

Finally, \mathcal{H}^n coincides with the Lebesgue measure \mathcal{L}^n up to a factor:

- (vi) If $A \subset \mathbb{R}^n$ a Borel set, then

$$\mathcal{H}^n(A) = \frac{1}{c_n} \mathcal{L}^n(A),$$

where c_n is the volume of the n -dimensional unit ball.

The Hausdorff measure behaves well under Lipschitz mappings, and more generally under Hölder mappings. in particular, this implies the scaling property (behavior under similarity transformations).

Theorem 5.11. *Let $F \subset \mathbb{R}^n$.*

- (i) *Let $f : F \rightarrow \mathbb{R}^m$ be a Hölder map, i.e., there exist $\alpha > 0$ and $c > 0$ such that*

$$|f(x) - f(y)| \leq c |x - y|^\alpha$$

for all $x, y \in F$. Then

$$\mathcal{H}^{\frac{s}{\alpha}}(f(F)) \leq c^{\frac{s}{\alpha}} \mathcal{H}^s(F)$$

for all $s \geq 0$.

- (ii) *Let $f : F \rightarrow \mathbb{R}^m$ be a Lipschitz map, i.e., there exist $c > 0$ such that*

$$|f(x) - f(y)| \leq c |x - y|$$

for all $x, y \in F$. Then

$$\mathcal{H}^s(f(F)) \leq c^s \mathcal{H}^s(F)$$

- (iii) *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a similarity transformation with scale factor $\lambda > 0$. Then*

$$\mathcal{H}^s(f(F)) = \lambda^s \mathcal{H}^s(F)$$

Proof.

- (i) Similar to proof of Theorem 5.7.
- (ii) Follows from (i) with $\alpha = 1$.
- (iii) Similarity transformations are bi-Lipschitz maps. Apply (ii) to f and f^{-1} .

□

For any $\delta < 1$, the function \mathcal{H}_δ^s is decreasing with $s \leq 0$. thus the function $\mathcal{H}^s = \lim_{\delta \rightarrow \infty} \mathcal{H}_\delta^s$ is decreasing with s . Even more, for a δ -cover $(U_i)_{i=1}^\infty$ of F , and $t > s$,

$$\sum_{i=1}^\infty \text{diam}(U_i)^t = \sum_{i=1}^\infty \text{diam}(U_i)^{t-s} \text{diam}(U_i)^s \leq \delta^{t-s} \sum_{i=1}^\infty \text{diam}(U_i)^s,$$

and taking infima over all δ -covers, we have

$$\mathcal{H}_\delta^t \leq \delta^{t-s} \mathcal{H}_\delta^s.$$

In the limit $\delta \rightarrow \infty$, we find that if $\mathcal{H}^s < \infty$, then $\mathcal{H}^t = 0$ for $t > s$. Thus, for a critical value $s = \dim_{\mathcal{H}} F$, the function \mathcal{H}^s jumps from ∞ to 0:

$$\mathcal{H}^s = \begin{cases} \infty & \text{if } s < \dim_{\mathcal{H}} F, \\ 0 & \text{if } s > \dim_{\mathcal{H}} F. \end{cases}$$

Formally, we define the Hausdorff dimension in the following way:

Definition 5.7. let $F \subset \mathbb{R}^n$, then

$$\dim_{\mathcal{H}} F = \inf \{s \geq 0 \mid \mathcal{H}^s = 0\} = \sup \{s \geq 0 \mid \mathcal{H}^s = \infty\}$$

is called the *Hausdorff dimension* of F .

Remark 5.11. At the critical value $s = \dim_{\mathcal{H}} F$, the function \mathcal{H}^s may be 0, ∞ , or

$$0 < \mathcal{H}^s < \infty.$$

Remark 5.12. The Hausdorff dimension can be defined without referring to the Hausdorff measure. to this end, let

$$\mathcal{H}_\infty^s(F) = \inf \left\{ \sum_{i=1}^\infty \text{diam}(U_i)^s \mid (U_i)_{i=1}^\infty \text{ is a countable cover of } F \right\}$$

Then

$$\dim_{\mathcal{H}} F = \inf \{s \geq 0 \mid \mathcal{H}_\infty^s = 0\}.$$

Remark 5.13. The Hausdorff measure and the Hausdorff dimension do not change if we restrict the covers to just open sets or just closed sets.

Similar to Theorem 5.5, one could also think of replacing the coverings involving arbitrary sets of diameter δ by say δ -balls, and consider

$$\mathcal{B}_\delta^s = \inf \left\{ \sum_{i=1}^\infty \text{diam}(U_i)^s \mid (U_i)_{i=1}^\infty \text{ is a cover of } F \text{ by } \delta\text{-balls} \right\} \geq \mathcal{H}_\delta^s$$

Then $\mathcal{B}^s = \lim_{\delta \rightarrow 0} \mathcal{B}_\delta^s$ leads to measure different from the Hausdorff measure, yet to the same dimension.

Considering packings by balls (“dual”) to coverings by balls) leads to packing measures and packing dimensions, both different from (yet considered to be closely related to) Hausdorff measures and Hausdorff dimension.

The Hausdorff dimension is always bounded from above by the lower box counting dimension.

Proposition 5.12. *Let $F \subset \mathbb{R}^n$ be bounded. Then*

$$\dim_H F \leq \underline{\dim}_B F.$$

Proof. Follows from $\mathcal{H}_\delta^s \leq N_\delta \delta^s$, where N_δ is the least number of sets of diameter δ that can cover F . \square

The behavior of the Hausdorff measure under Hölder and thus Lipschitz mappings (Theorem 5.11), implies corresponding properties for the Hausdorff dimension.

Theorem 5.13.

(i) *Let $f : F \rightarrow \mathbb{R}^m$ be a Hölder map, i.e., there exist $\alpha > 0$ and $c > 0$ such that*

$$|f(x) - f(y)| \leq c|x - y|^\alpha$$

for all $x, y \in F$. Then

$$\dim_H f(F) \leq \frac{1}{\alpha} \dim_H F$$

(ii) *Let $f : F \rightarrow \mathbb{R}^m$ be a Lipschitz map, i.e., there exist $c > 0$ such that*

$$|f(x) - f(y)| \leq c|x - y|$$

for all $x, y \in F$. Then

$$\dim_H f(F) \leq \dim_H F$$

(iii) *If $f : F \rightarrow \mathbb{R}^m$ is a bi-Lipschitz map, i.e., there exist constants $0 < c_1 \leq c_2$ such that*

$$c_1|x - y| \leq |f(x) - f(y)| \leq c_2|x - y|$$

for all $x, y \in F$. Then

$$\dim_H f(F) = \dim_H F.$$

Proof.

(i) By Theorem 5.11, for $s > \dim_H F$

$$\mathcal{H}_\alpha^s(f(F)) \leq c^{\frac{s}{\alpha}} \mathcal{H}^s(F) = 0.$$

Thus, $\dim_H f(F) \leq \frac{s}{\alpha}$ for all $s > \dim_H F$.

(ii) Follows from (i) with $\alpha = 1$.

(iii) Apply (ii) to f and f^{-1} .

\square

Remark 5.14. Similar to Remark 5.8, we obtain that the Hausdorff dimension can only reduce under projection, and is invariant under affine transformations.

Remark 5.15. One approach to fractal geometry is to regard two sets as equivalent if there exists a bi-Lipschitz map between them. Since bi-Lipschitz maps are homeomorphisms, topological invariants are bi-Lipschitz invariants, while Hausdorff dimension provides a further invariant to distinguish equivalence classes.

Vice versa, the Hausdorff dimension provides little information about the topology of a set. However, every set $F \subset \mathbb{R}^n$ with $\dim_H F < 1$ is totally disconnected (no two points lie in the same connected component).

Again we summarize properties of the Hausdorff dimension, most of which follow directly from the properties of Hausdorff measures.

Theorem 5.14. *The Hausdorff dimension has the following properties:*

(i) **Monotonicity.** *If $E \subset F$, then*

$$\dim_H E \leq \dim_H F.$$

(ii) **Range of values.**

$$0 \leq \dim_H F \leq n.$$

(iii) **Finite stability.** *\dim_H is finitely stable, i.e.,*

$$\dim_H(E \cup F) = \max\{\dim_H E, \dim_H F\}.$$

(iv) **Countable stability.** $\dim_H(\bigcup_{i=1}^{\infty} F_i) = \sup_i \dim_H F_i$.

(v) **Finite sets.** $\dim_H F = 0$ if F is a finite set.

(vi) **Countable sets.** $\dim_H F = 0$ if F is countable.

(vii) **Open sets.** $\dim_H F = n$ if F is an open subset of \mathbb{R}^n .

(viii) **Smooth manifold.** $\dim_H F = d$ if F is a smooth d -dimensional submanifold of \mathbb{R}^n .

(ix) **Geometric invariance.** See Remark 5.14.

(x) **Lipschitz invariance.** See Theorem 5.13.

5.2.3 Similarity dimension

We had seen that for a self-similar set, applying a measure satisfying the scaling property, leads to the following reasonable definition of dimension.

Definition 5.8. Let $F \subset \mathbb{R}^n$ be a self-similar set, i.e., the attractor of an IFS $\{S_1, \dots, S_m\}$ of similarity transformations with scaling factors $0 < r_i < 1$, $i = 1, \dots, m$. Then the number

$$\dim_S F = s$$

with

$$\sum_{i=1}^m r_i^s = 1$$

is called the *similarity dimension* of F .

By the heuristic argument made earlier we have also seen that it should coincide with the Hausdorff dimension if the union in

$$F = \bigcup_{i=1}^m S_i(F)$$

is disjoint. It even holds if the sets $S_i(F)$ do not overlap “too much”, which is described by an “open set condition” as a condition on the similarity transformations S_i . In this case the similarity dimension also coincides with the box-counting dimension.

Theorem 5.15. *Let $\{S_1, \dots, S_m\}$ be an IFS of similarity transformations with scaling factors $0 < r_i < 1$, $i = 1, \dots, m$ satisfying the **open set condition**, i.e., there exists a non-empty bounded open set U such that*

$$U \supset \bigcup_{i=1}^m S_i(U).$$

Then the attractor

$$F = \bigcup_{i=1}^m S_i(F)$$

of the IFS satisfies

$$\dim_H F = \dim_B F = \dim_S F.$$

Moreover, for this value $s = \dim_S F$ the Hausdorff measure satisfies

$$0 < \mathcal{H}^s < \infty.$$

Remark 5.16. In the special case $r = r_i$, $i = 1, \dots, m$, one obtains

$$\dim_S F = \frac{\log m}{-\log r}.$$

Remark 5.17. In particular, if $S_1(F), \dots, S_m(F)$ are disjoint, the open set condition holds.

Example 5.6. For the Sierpinski triangle, we have $r = \frac{1}{2}$, $m = 3$, and thus,

$$\dim_S F = \frac{\log 3}{\log 2}.$$

Remark 5.18. Not assuming the open set condition, it still holds that

$$\dim_H F = \dim_B F \leq \dim_S F.$$

5.3 Iteration of complex functions

5.3.1 Julia sets

^{JT}: [The theory for Julia sets is almost the same for rational functions, provided that ∞ is included in the natural way. The main difference is that $J(f)$ may not be bounded (still closed), and it may have interior points (in which case $J = \mathbb{C} \cup \{\infty\}$).

A motivation can be analysing the Newton method.]

For a complex function $f : \mathbb{C} \rightarrow \mathbb{C}$ we investigate the behavior of the sequences

$$(f^k(z))_{k=0}^{\infty} = \left(\underbrace{f \circ \cdots \circ f}_{k \text{ times}}(z) \right)_{k=0}^{\infty}$$

for different $z \in \mathbb{C}$. The *Julia set* will be the curve that is the interface between different qualitative behaviors of that sequence.

Example 5.7.

(i) Consider the function

$$f(z) = z^2,$$

so that $f^k(z) = z^{2^k}$. For a point $z \in \mathbb{C}$ with $|z| < 1$ the sequence $f^k(z)$ converges to the origin

$$f^k(z) \rightarrow 0 \quad (k \rightarrow \infty).$$

For a point $z \in \mathbb{C}$ with $|z| > 1$ the (absolute value of the) sequence $f^k(z)$ converges to infinity

$$f^k(z) \rightarrow \infty \quad (k \rightarrow \infty).$$

And for a point $z \in \mathbb{C}$ with $|z| = 1$ the sequence remains bounded, staying on the unit circle. Thus, the unit circle \mathbb{S}^1 is interface between the two different behaviors of converging to ∞ and converging to 0, or more generally, staying bounded. The unit circle is the Julia set of f in this example.

(ii) Consider the function

$$f(z) = z^2 + c.$$

for some $c \in \mathbb{C}$ with small absolute value. Then for sufficiently small $|z|$ the sequence $f^k(z)$ converges to the fixed point of f close to 0, and for sufficiently large $|z|$ the sequence $f^k(z)$ converges to infinity. In between there is a curve that is the interface between these two behaviors, which in this case will be a fractal.

We will restrict our following considerations to polynomial functions f .

Definition 5.9. Let $f : \mathbb{C} \rightarrow \mathbb{C}$ be a polynomial function of degree $n \geq 2$, i.e.,

$$f(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0$$

with some complex numbers $a_k \in \mathbb{C}$, $k = 0, \dots, n$, $a_n \neq 0$.

(i) The *filled-in Julia set* of f is given by

$$K(f) = \left\{ z \in \mathbb{C} \mid \lim_{k \rightarrow \infty} f^k(z_1) \neq \infty \right\}.$$

(ii) The *Julia set* of f is given by the boundary of $K(f)$

$$J(f) = \partial K(f)$$

i.e., $z \in J(f)$ if in every neighborhood of z there are points z_1 such that $\lim_{k \rightarrow \infty} f^k(z_1) = \infty$ and points z_2 such that $\lim_{k \rightarrow \infty} f^k(z_2) \neq \infty$.

(iii) The *Fatou set* (or *stable set*) of f is the complement of the Julia set

$$F(f) = \mathbb{C} \setminus J(f).$$

Remark 5.19. The definition of the filled-in Julia K set suggests the following way of visualization (we will encounter other methods later): Let $r > 0$ be large, and $N \in \mathbb{N}$ large. For every point $z \in \mathbb{C}$ (e.g. on some grid) compute successive terms of the sequence $(f^k(z))_{k=0}^{\infty}$ until

- $|f^k(z)| \geq r$, then we consider z to be not in K
- or $k = N$, in which case we consider z to be contained in K .

To only (or additionally) draw the Julia set J , for every point in your grid, check the behavior of $f^k(z)$ with the method before. If all four corners show the same behavior consider z to be not in J , and other wise consider z to be in J .

For a polynomial the sequence $f^k(z)$ converges to ∞ if and only if it is unbounded as the following lemma ensures. ^{JT:} [Follows directly from Liouville's theorem.]

Lemma 5.16. *Let f be a polynomial of degree $n \geq 2$.*

(i) *There exists an $r > 0$ such that if $z \in \mathbb{C}$ with $|z| \geq r$ then*

$$|f(z)| \geq 2|z|.$$

(ii) *There exists an $r > 0$ such that if $z \in \mathbb{C}$ with $|f^m(z)| \geq r$ for some $m \in \mathbb{N}$ then*

$$\lim_{k \rightarrow \infty} f^k(z) = \infty \quad (k \rightarrow \infty)$$

(iii) *For any $z \in \mathbb{C}$, either $(f^k(z))_{k=0}^{\infty}$ is bounded or $\lim_{k \rightarrow \infty} f^k(z) = \infty$.*

Proof.

(i) For $z \in \mathbb{C}$ we have

$$|f(z)| \geq |a_n| |z|^n - (|a_{n-1}| |z|^{n-1} + \cdots + |a_0|).$$

Choose $r > 0$ such that for $z \in \mathbb{C}$ with $|z| > r$ we have

$$\frac{1}{2} |a_n| |z|^n \geq |a_{n-1}| |z|^{n-1} + \cdots + |a_0| \quad \text{and} \quad \frac{1}{2} |a_n| |z|^n \geq 2|z|.$$

Then

$$|f(z)| \geq \frac{1}{2} |a_n| |z|^n \geq 2|z|.$$

(ii) If $|f^m(z)| \geq r$, then

$$|f^{m+k}(z)| \geq 2^k |f^m(z)| \geq 2^k r \rightarrow \infty \quad (k \rightarrow \infty).$$

(iii) If the sequence is unbounded, it will eventually satisfy the condition in (ii).

□

Thus the Julia set is the interface between the points for which the sequence $f^k(z)$ is unbounded and those for which it is bounded. The following proposition immediately follows.

Proposition 5.17. $J(f^p) = J(f)$ for every positive integer p .

Proof. The condition in Lemma 5.16 (ii) is satisfied for $(f^k(z))_{k=0}^\infty$ if and only if it is satisfied for $((f^p)^k(z) = (f^{pk}(z))_{k=0}^\infty$. \square

Furthermore, we can use Lemma 5.16 to infer some basic topological features of the Julia set.

Proposition 5.18. *The Julia set $J(f)$ and the filled-in Julia set $K(f)$ are both non-empty and compact with $J(f) \subset K(f)$. Furthermore, $J(f)$ has an empty interior.*

Proof. By Lemma 5.16, both $K(f)$ and its boundary $J(f)$ must be bounded.

We show that the complement of $K(f)$ is open. Let $z \notin K(f)$. Then $\lim_{k \rightarrow \infty} f^k(z) = \infty$ and $|f^m(z)| > r$ for some integer m . By continuity of f^m this still holds for w in a small neighborhood of z . By Lemma 5.16, $\lim_{k \rightarrow \infty} f^k(w) = \infty$, and thus $w \notin K(f)$. Thus, $K(f)$ is closed.

$J(f)$ is the boundary of $K(f)$ and thus closed. Since $K(f)$ is closed, we also have $J(f) \subset K(f)$.

To see that $K(f)$ is not empty, let z_0 be a solution to the equation $f(z) = z$ (there exists at least one). Then $f^k(z_0) = z_0$ and thus $z_0 \in K(f)$. On the other hand, by Lemma 5.16, the complement $\mathbb{C} \setminus K(f)$ is not empty. Thus, let $z_1 \in \mathbb{C} \setminus K(f)$. The line segment connecting z_0 and z_1 must have at least one point on the boundary $J(f)$. Thus, $J(f)$ is not empty.

The boundary of any set has empty interior. \square

The Julia set is invariant under the map f .

Proposition 5.19. *The Julia set $J = J(f)$ is forward and backward invariant under the map f , i.e.,*

$$f(J) = J = f^{-1}(J).$$

Proof. Let $z \in J \subset K$. Then may find a sequence $(z_n)_{n=0}^\infty \subset \mathbb{C} \setminus K$ with $\lim_{n \rightarrow \infty} z_n = z$. Thus, we have

$$\lim_{n \rightarrow \infty} f^k(z) \neq \infty \quad \text{and} \quad \lim_{k \rightarrow \infty} f^k(z_n) = \infty.$$

and therefore

$$\lim_{n \rightarrow \infty} f^k(f(z)) \neq \infty \quad \text{and} \quad \lim_{k \rightarrow \infty} f^k(f(z_n)) = \infty.$$

By continuity of f , we can choose $f(z_n)$ arbitrarily close to $f(z)$, and thus $f(z) \in J$. So, we have

$$J \subset f(J) \quad \text{and therefore also} \quad J \subset f^{-1}(f(J)) \subset f^{-1}(J).$$

With z and $(z_n)_{n=0}^\infty$ as before, let w such that $f(w) = z$. Then since f is a polynomial, we may find $(w_n)_{n=0}^\infty$ with $\lim_{n \rightarrow \infty} w_n = w$ and $f(w_n) = z_n$. Thus

$$\lim_{n \rightarrow \infty} \underbrace{f^k(w)}_{=f^{k-1}(z)} \neq \infty \quad \text{and} \quad \lim_{k \rightarrow \infty} \underbrace{f^k(w_n)}_{=f^{k-1}(z_n)} = \infty,$$

and therefore $w \in J$. So, we have

$$f^{-1}(J) \subset J \quad \text{and therefore also} \quad J = f(f^{-1}(J)) \subset f(J).$$

\square

The following theorem is a consequence of Montel's theorem on normal sequences of holomorphic functions. ^{JT}: [What does (the existence of) the exceptional point say about f ?]

Theorem 5.20. *Let $z \in J(f)$ and U a neighborhood of z . Then the set*

$$\bigcup_{k=l}^{\infty} f^k(U), \quad l = 0, 1, \dots$$

is the whole of \mathbb{C} , except for possibly one single point. Any such point is called an exceptional point, is not contained in $J(f)$, and is independent of z and U .

Proof. By use of Montel's theorem. □

Remark 5.20. To use Montel's theorem one may show that Julia sets may equivalently be represented in the following way:

$$J(f) = \{z \in \mathbb{C} \mid (f^k)_{k=1}^{\infty} \text{ is normal at } z\}.$$

As a consequence, we prove the following theorem, which may be used for visualization of Julia sets.

Theorem 5.21.

(i) *The following holds for all $z \in \mathbb{C}$ except at most one exception: Let U be an open set with*

$$U \cap J(f) \neq \emptyset.$$

Then $f^{-k}(z)$ intersects U for infinitely many values of k . If there is an exceptional value, it cannot be in $J(f)$.

(ii) *For any $z \in J(f)$*

$$J(f) = \overline{\bigcup_{k=0}^{\infty} f^{-k}(z)}.$$

Proof.

(i) By Theorem 5.20, a non-exceptional point $z \in U$ satisfies $z \in f^k(U)$ for some k , and thus $f^{-k}(z)$ intersects U . Applying this repeatedly, we can generate arbitrary large k for which this holds.

(ii) Let $z \in J(f)$.

By Proposition 5.19, this implies $f^{-k}(z) \subset J(f)$, and thus

$$\bigcup_{k=0}^{\infty} f^{-k}(z) \subset J(f),$$

since $J(f)$ is closed.

Recall that a point z is in the closure of a set A if every neighborhood of z contains points of A . By (i), for a neighborhood U that intersects $J(f)$ (which is a neighborhood of a point in $J(f)$) there exists a k such that $f^{-k}(z)$ intersects U . Thus

$$J(f) \subset \overline{\bigcup_{k=0}^{\infty} f^{-k}(z)}.$$

□

We may use this to show that the Julia set has no isolated points.

Proposition 5.22. *$J(f)$ is a perfect set (closed and with no isolated points) and therefore uncountable.*

Proof. Let $w \in J(f)$ and U a neighborhood of w . We have to show that U contains points of $J(f)$ other than w . By Theorem 5.21, U contains a point of $f^{-k}(w) \subset J(f)$ for some $k \geq 1$. If this point is different from w , we are done. However, this point might not be different from w . It is equal to w if and only if

$$f^k(w) = w.$$

Consider the case $k = 1$ first (w a fixed point of f). Thus, $f(w) = w$. If there is no other solution to the equation $f(z) = w \in J(f)$, this would contradict Theorem 5.21 (ii). ^{JT:} [Unless J consists of exactly one point. Is that possible?] Thus, let $v \neq w$ be such that $f(v) = w$. Again, by Theorem 5.21, U contains a point of $f^{-l}(v) \subset f^{-l-1}(w) \subset J(f)$ for some $l \geq 1$. Any such point u is distinct from w , since $f^l(u) = v \neq w = f^l(w)$.

Now assume $k > 1$ (w is a periodic orbit of f). Thus, $f^k(w) = w$ and w is a fixed point of f^k . By Theorem 5.17, $J(f^k) = J(f)$, and we may apply the previous argument to f^k . ^{JT:} [one could just earlier consider $\tilde{f} = f^k$, so that the order of arguments doesn't have to be reversed.] □

In the proof of the previous proposition, we have encountered fixed points and periodic orbits of f , which may be used to characterize Julia sets.

Definition 5.10. Let $f : \mathbb{C} \supset U \rightarrow \mathbb{C}$ be a holomorphic function, and $w \in U$.

- (i) If $f(w) = w$, the point w is called a *fixed point* of f .
In this case w is called *attractive* if $|f'(w)| < 1$ and it is called a *repelling* if $|f'(w)| > 1$.
- (ii) If $f^p(w) = w$ for some $p \geq 1$, the point w is called a *periodic point* of f .
The smallest such p is called the *period* of w , and $w, f(w), \dots, f^{p-1}(w)$ a *periodic orbit* of f . w is called *attractive* if $|(f^p)'(w)| < 1$ and it is called *repelling* if $|(f^p)'(w)| > 1$.

Remark 5.21. Close to a fixed point w the function f may be expressed as

$$f(z) = f(w) + f'(w)(z - w) + o(z - w) = w + f'(w)(z - w) + o(z - w)$$

It locally acts as a similarity transformation centered at w with scaling factor $|f'(w)|$. Thus, if $|f'(w)| < 1$, points close to w get closer to w after applying f and thus get attracted in the sequence $f^k(w)$. If $|f'(w)| > 1$, points close to w get repelled from w .

For a periodic point w of period p we consider the same expression for

$$f^p(w) = w + (f^p)'(w)(z - w) + o(z - w)$$

Thus, if $|(f^p)'(w)| < 1$, points close to w , get closer to w after further p applications of f and thus get attracted to the periodic orbit. If $|(f^p)'(w)| > 1$, points close to w get repelled from the periodic orbit.

Note that by successive application of the chain rule, the derivative $(f^p)'(w)$ may be expressed as

$$(f^p)'(w) = f'(f^{p-1}(w)) \cdot f'(f^{p-2}(w)) \cdots f'(w).$$

Thus, its value does not depend on the point in the periodic orbit.

For later reference, we state the following lemma on attractive periodic orbits.

Lemma 5.23. *Let f be a polynomial of degree $n \geq 2$, and $w(\neq \infty)$ be an attractive periodic point. Then there exists a $z \in \mathbb{C}$ with $f'(z) = 0$ such that the sequence $(f^k(z))_{k=0}^{\infty}$ is attracted to the periodic orbit of w .*

In particular this means, that there can be at most as many attractive periodic orbits as there are critical points ($f'(z) = 0$), which in turn are at most $n - 1$.

On the other hand, f has a large number of repelling periodic points. In fact their closure constitute the entire Julia set.

Theorem 5.24. *Let f be a polynomial of degree $n \geq 2$. Then $J(f)$ is the closure of the repelling periodic points of f .*

Proof. By use of Montel's theorem once more. □

Remark 5.22. Some additional properties of $J(f)$ related to the previous theorem are the following.

- ▶ Periodic orbits are dense in $J(f)$. However, there are also points $z \in J(f)$ such that the sequence $f^k(z)$ is dense in $J(f)$.
- ▶ The dependence of f on initial conditions (starting value $z \in J(f)$) is sensitive on $J(f)$. Thus, distances $|f^k(z) - f^k(w)|$ can become large no matter how close z and w are chosen.
- ▶ This may be summarized as “ f acts chaotically on $J(f)$ ”.

Another way to characterize Julia sets, is as the boundary of the basin of attractive fixed points.

Definition 5.11. Let $w \in \mathbb{C} \cup \{\infty\}$ be an attractive fixed point of f . Then

$$A(w) = \left\{ z \in \mathbb{C} \mid \lim_{k \rightarrow \infty} f^k(z) = w \right\}$$

is called the *basin of attraction* of w .

Remark 5.23. Note that by Lemma 5.16, ∞ is always an attractive fixed point of a polynomial f .

Lemma 5.25. *The basin of attraction $A(w)$ is open.*

Proof. Since w is attractive there is an open set $U \subset A(w)$. Thus, $f^{-k}(U) \subset A(w)$ is open for all $k \geq 0$. For $z \in A(w)$, we have $f^k(z) \in U$ for some $k \geq 0$ and thus $z \in f^{-k}(U)$. □

Theorem 5.26. *Let f be a polynomial of degree $n \geq 2$, and $w \in \mathbb{C}$ be an attractive fixed point of f . Then*

$$J(f) = \partial A(w).$$

Proof. Let $z \in J(f)$. Then $f^k(z) \in J(f)$ for all k , and thus $z \notin A(w)$. However, if U is a neighborhood of z , by Theorem 5.20, the set $f^k(U)$ contains point of $A(w)$ for some k . Thus, there are points arbitrary close to z that iterate to w , and therefore $z \in \overline{A(w)}$. So,

$$J(f) \subset \partial A(w).$$

The reverse inclusion may be shown by use of Montel's theorem once more. □

5.3.2 Julia sets of quadratic polynomials (and the Mandelbrot set)

We now restrict to our study of Julia sets to quadratic polynomials.

Lemma 5.27. *Let*

$$f(z) = a_2 z^2 + a_1 z + a_0$$

be a quadratic polynomial, $a_2 \neq 0$. Then there exists a complex linear function $h(z) = \alpha z + \beta$, $\alpha \neq 0$ such that

$$f_c(z) := h \circ f \circ h^{-1}(z) = z^2 + c$$

for some $c \in \mathbb{C}$.

Proof. We go the reverse direction and substitute

$$h^{-1} \circ f_c \circ h(z) = \alpha z^2 + 2\beta z + \frac{\beta^2 + c - \beta}{\alpha}.$$

From this arbitrary values for a_2, a_1, a_0 may be obtained. □

Geometrically, the linear function h is a similarity transformation. Thus, we may restrict our study further to the functions

$$f_c(z) = z^2 + c$$

with some $c \in \mathbb{C}$. The Julia sets will be similar to the corresponding ones obtained from general quadratic polynomials.

We gather some facts that let us better understand how $f_c : \mathbb{C} \rightarrow \mathbb{C}$ acts as a holomorphic function.

- f_c is locally bijective away from every point except the critical point

$$z = 0, \quad f'_c(0) = 0.$$

In fact, it is bijective on every open half-plane, whose boundary contains 0, which is mapped by f_c to a slitted plane, where the slit is a ray starting from c .

- The preimage of any point $z \neq 0$ consists of two points, the two square roots

$$f_c^{-1}(z) = \pm(z - c)^{\frac{1}{2}}$$

The two branches of the square roots, may be considered as the two holomorphic functions which are inverse to f_c on two complementary half-planes.

- We call a smooth, closed, simple (non-self-intersecting) curve in the complex plane a *loop*. It is the boundary of a simply-connected set, which is the inside of the loop. A loop winds around every point of its interior exactly once. Let $L = \partial U$ be a loop in the complex plane, and U its simply-connected interior.

If $0 \in U$, then $f_c(L)$ is a loop.

If $0 \notin U$, then $f_c(L)$ is a curve that wraps around c twice.

The following lemma states what the preimages of loops look like.

Lemma 5.28. *Let $L = \partial U \subset \mathbb{C}$ be a loop in the complex plane and U its simply-connected interior.*

(i) If $c \in U$, then $f^{-1}(L)$ is a loop.

Moreover, f_c maps the interior of $f_c^{-1}(L)$ onto the interior of L , and the exterior of $f_c^{-1}(L)$ onto the exterior of L .

(ii) If $c \notin U$, then $f^{-1}(L)$ comprises of two disjoint loops, neither contained within the other.

Moreover, f_c maps the interior of each loop of $f_c^{-1}(L)$ onto the interior of L , and the region outside both loops of $f_c^{-1}(L)$ onto the exterior of L .

Remark 5.24. If $c \in L$, then $f_c^{-1}(L)$ is a “figure of eight” (a single point of self-intersection).

Consider a large loop $L = \partial U$ such that $c \in U$. Then Theorem 5.21 suggests, that the sequence $(f^{-k}(L))_{k=0}^{\infty}$ should converge to the Julia set.

Then by Lemma 5.28, each curve in the sequence $(f^{-k}(L))_{k=0}^{\infty}$ is a single loop as long as c lies inside the previous loop, or equivalently, as long as $(f^k(0))_{k=0}^{\infty} \in U$.

Thus, we may distinguish two cases. One in which the sequence $(f^k(0))_{k=0}^{\infty}$ stays bounded, and the sequence $(f^{-k}(L))_{k=0}^{\infty}$ consists of single loops. In this case, the limit set, which is the Julia set, is connected. And the other case, in which the sequence $(f^k(0))_{k=0}^{\infty}$ goes to infinity, and the sequence $(f^{-k}(L))_{k=0}^{\infty}$ at some point disconnects into multiple components. In this case, the Julia set is disconnected, in fact even totally disconnected.

Theorem 5.29. *Let $c \in \mathbb{C}$. Then $(f^k(0))_{k=0}^{\infty}$ is bounded if and only if $J(f_c)$ is connected. Furthermore, if $J(f_c)$ is not connected, then it is totally disconnected.*

Proof.

(\Rightarrow) Let $(f_c^k(0))_{k=0}^{\infty}$ be bounded. Let $L = \partial U \subset \mathbb{C}$ be a large circle with interior U , such that

(i) $(f_c^k(0))_{k=0}^{\infty} \subset U$,

(ii) $f_c^{-1}(L) \subset U$,

(iii) and all points outside L iterate to ∞ .

By (i) and Lemma 5.28, the sequence $(f_c^{-k}(L))_{k=0}^{\infty}$ consists of single loops. By (ii) $f_c^{-2}(U) \subset f_c^{-1}(U)$. Thus, $(f_c^{-k}(L))_{k=0}^{\infty}$ is a sequence of loops, each in the interior (or on) of the previous one.

Let K be the closed set points inside or on every of the loops $f_c^{-k}(L)$:

$$K = \bigcap_{k=0}^{\infty} f_c^{-k}(\overline{U})$$

For $z \in K$ all points $f_c^k(z) \in U$, and thus $(f_c^k(z))_{k=0}^{\infty}$ is bounded. If $z \notin K$, then z is outside one of the loops $f_c^{-m}(L)$ for some m , and therefore, $f_c^m(z) \notin U$. By (iii), $\lim_{k \rightarrow \infty} f_c^k(z)$. Thus, K is the filled-in Julia set of f_c .

The sequence $f_c^{-k}(\overline{U})$ is a decreasing sequence of closed simply-connected sets. Thus, K is simply-connected and therefore has connected boundary $\partial K = J(f_c)$.

(\Leftarrow) Let $(f_c^k(0))_{k=0}^{\infty}$ be unbounded. Let $L = \partial U \subset \mathbb{C}$ be a large circle with interior U , such that

- (i) $(f_c^k(0))_{k=0}^\infty \not\subset L$,
- (ii) $f_c^{-1}(L) \subset U$,
- (iii) and all points outside L iterate to ∞ .

Let m be the smallest integer such that $f_c^m(0)$ lies outside L .

Then $(f_c^{-k}(L))_{k=0}^{m-1}$ is a (finite) sequence of loops, each in the interior (or on) of the previous one. However, c is outside of the loop $f_c^{-m+1}(L)$. Thus, by Lemma 5.28, $f_c^{-m}(L)$ consists of two loops inside $f_c^{-m+1}(L)$, and f_c maps the interior of those loops onto the interior of $f_c^{-m+1}(L)$.

The Julia set $J(f_c)$ lies inside these two loops, since points outside iterate to infinity, by (iii). Furthermore, since $J(f_c)$ is invariant under f_c^{-1} , both loops must contain part of the Julia set, and thus, $J(f_c)$ is not connected.

If we continuously apply Lemma 5.28 to the sequence $(f_c^{-k}(L))_{k=0}^{m-1}$, we see that the Julia set lies within a “Cantor-like” hierarchy of disjoint pairs of loops, and therefore must be totally disconnected.

□

We now define the Mandelbrot set as all the values $c \in \mathbb{C}$ for which the Julia set $J(f_c)$ is connected, or equivalently, for which the sequence $(f_c^k(0))_{k=0}^\infty$ is bounded.

Definition 5.12. The *Mandelbrot set* is defined by

$$\begin{aligned} M &= \{c \in \mathbb{C} \mid J(f_c) \text{ is connected}\} \\ &= \left\{c \in \mathbb{C} \mid (f_c^k(0))_{k=0}^\infty \text{ is bounded}\right\} \\ &= \left\{c \in \mathbb{C} \mid \lim_{k \rightarrow \infty} f_c^k(0) \neq \infty\right\} \end{aligned}$$

Remark 5.25. The boundary of the Mandelbrot set has zero area, yet is a fractal of Hausdorff dimension $\dim_H M = 2$.

Remark 5.26. The first equality holds by Theorem 5.29 and the last by Lemma 5.16.

The characterization by Theorem 5.29 gives immediate rise to a way of visualizing the Mandelbrot set, similar to the method discussed in Remark 5.19: Choose $r > 2$ and $N \in \mathbb{N}$ large. For each $c \in \mathbb{C}$ compute successive terms of the sequence $(f_c^k(0))_{k=0}^\infty$ until

- $|f_c^k(0)| \geq r$, then $c \notin M$,
- or $k = N$, in which case we consider c to be contained in M .

One can additionally assign different colors to the complement of M , depending on the first number k for which $|f_c^k(0)| \geq r$.

Remark 5.27. The Mandelbrot set is bounded. At least the following bound may be given: For

$$|c| > \frac{1}{4}(5 + 2\sqrt{6}) \approx 2.475... \quad \Rightarrow \quad c \notin M,$$

and thus f_c is totally disconnected.

In this case the Julia set $J(f_c)$ is the attractor of the iterated function system (of contractions) consisting of the two branches $f_c^{-1}(z) = \pm(z - c)^{\frac{1}{2}}$ of the square root for z close to $J(f_c)$.

Furthermore, for large $|c|$, its dimension behaves like

$$\dim_{\text{B}} J(f_c) = \dim_{\text{H}} J(f_c) \sim \frac{2 \log 2}{\log(4|c|)}.$$

A finer distinction of different qualitative behavior of the Julia sets $J(f_c)$ than by the Mandelbrot set (and its complement), may be obtained by considering the attractive periodic orbits of f_c .

Lemma 5.30.

- (i) f_c has at most one attractive periodic orbit.
- (ii) If $c \notin M$ then f_c has no attractive periodic orbit

Proof.

- (i) By Lemma 5.23, for every attractive periodic orbit, there exists a critical point of f_c that is attracted to it. The only critical point of f_c is given by

$$f'_c(z) = 2z = 0$$

is given by $z = 0$. Thus, f_c has at most one attractive periodic orbit (including an attractive fixed point different from ∞).

- (ii) If $c \notin M$, then $\lim_{k \rightarrow \infty} f_c^k(0) = \infty$, and thus, by (i), f_c has no attractive periodic orbit.

□

The complement of the Mandelbrot set only contains $c \in M$ for which f_c has no attractive periodic points. It is conjectured that points $c \in M$ for which f_c does have attractive periodic points fill the interior of M .

Conjecture 5.31 (density of hyperbolicity). *The interior of M consists of the points $c \in \mathbb{C}$ for which f_c has an attractive periodic point.*

By Lemma 5.30, we have

$$f_c \text{ has attractive periodic point} \quad \Rightarrow \quad c \in M.$$

Now different arias inside the Mandelbrot set may be identified by the period p of the attractive orbit.

p = 1 f_c has an attractive fixed point, i.e.,

$$f_c(z) = z \quad \text{and} \quad |f_c(z)| < 1.$$

It can be shown that this is the case if and only if c lies inside a cardioid, called the “main cardioid” of the Mandelbrot set.

Theorem 5.32. *f_c has an attractive fixed point if and only if c lies inside the cardioid*

$$z(t) = \frac{1}{2}e^{it} \left(1 - \frac{1}{2}e^{it} \right), \quad t \in [0, 2\pi].$$

This is the cardioid obtained by rolling a circle of radius $\frac{1}{4}$ along another circle of radius $\frac{1}{4}$ and center 0 following the initial point of contact $\frac{1}{4}$.

Furthermore, in this case, $J(f_c)$ is a simple close loop.

Remark 5.28. For small $|c|$ the dimension of $J(f_c)$ behaves like

$$\dim_{\text{B}} J(f_c) = \dim_{\text{H}} J(f_c) = 1 + \frac{|c|^2}{4 \log 2} + O(|c|^3)$$

Moreover, $0 < \mathcal{H}^s < \infty$.

p=2 f_c has an attractive periodic orbit of period 2, i.e.

$$f_c^2(z) = z \quad \text{and} \quad |(f_c^2)'(z)| < 1.$$

Lemma 5.33. f_c has an attractive fixed point if and only if c lies in the disk

$$|c + 1| < \frac{1}{4}$$

f_c has two fixed points and two period 2 orbits (since f_c^2 has degree 4), one of which is attractive. Let w_1 and w_2 be the two points of the attractive period 2 orbit. Both points are fixed points of f_c^2 , thus by Theorem 5.26 and Proposition 5.17

$$J(f_c) = J(f_c^2) = \partial A(w_1, f_c^2) = \partial A(w_2, f_c^2).$$

It turns out that the region of the basins of attraction containing w_1 and w_2 each are bounded by a simple closed curve, which touch each other at a fixed point of f_c . The Julia set consists of all preimages of these two loops, enclosing all preimages of w_1 and w_2 , and always touching each other pairwise in “pinch points”.

p > 2 The Julia set consists of all preimages of p loops, each enclosing one of the points w_1, \dots, w_p of the attractive period p orbit. The preimages of these loops enclose all the preimages of the points w_1, \dots, w_p and touch each other in tuples of p .

boundary There exist more intricate Julia sets $J(f_c)$ for values on the boundary of M .

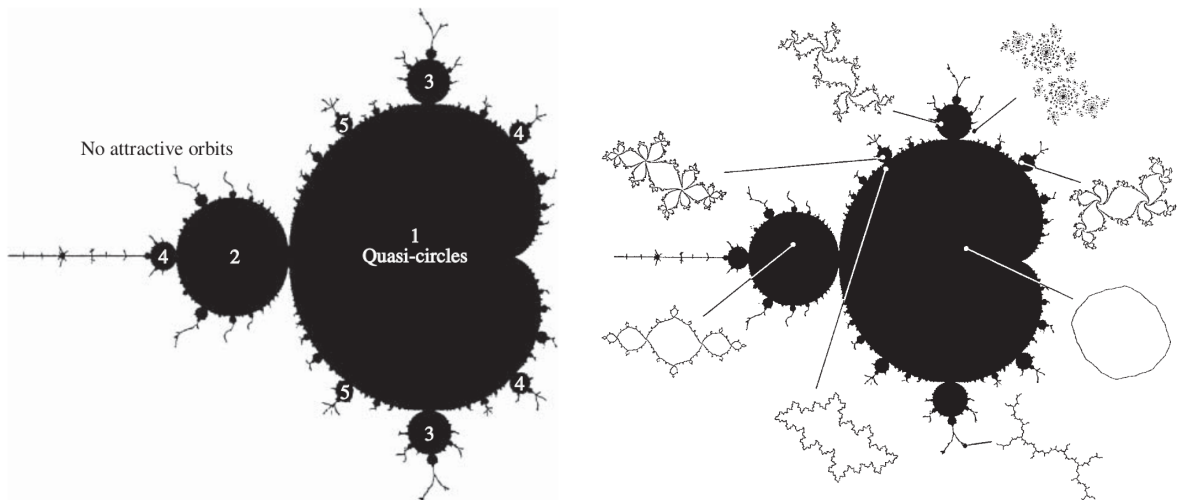


Figure 18. Different regions of the Mandolbrot set according to the period of attractive periodic orbits and the corresponding Julia sets.

6 Möbius geometry

6.1 The elementary model of Möbius geometry

Consider the n -dimensional Euclidean space \mathbb{R}^n . The inversion in a hypersphere with center $c \in \mathbb{R}^n$ and radius $r > 0$ can be described in the following way: The point x and its image x' lie on the same ray emanating from c and the distances to c satisfy the relation

$$\|x - c\| \cdot \|x' - c\| = r^2.$$

This gives rise to an involution on \mathbb{R}^n , except that the center c has no image and no preimage. To fix this, we add one extra point to \mathbb{R}^n , called ∞ , and obtain the *extended Euclidean space*

$$\widehat{\mathbb{R}^n} := \mathbb{R}^n \cup \{\infty\}.$$

Definition 6.1. The (*sphere*) *inversion* in the hypersphere with center $c \in \mathbb{R}^n$ and radius $r > 0$ is the map defined by

$$\begin{aligned} \widehat{\mathbb{R}^n} &\rightarrow \widehat{\mathbb{R}^n}, & x &\mapsto x' = c + \frac{r^2}{\|x - c\|^2} (x - c) && \text{for } x \neq c, \\ & & c &\mapsto \infty \\ & & \infty &\mapsto c \end{aligned}$$

Sphere inversions preserve angles and map hyperspheres and hyperplanes to hyperspheres and hyperplanes. This statement becomes simpler and more specific at the same time if we consider hyperplanes as special cases of hyperspheres through the point ∞ . More precisely, let us adopt the following convention:

Definition 6.2. A *sphere* in $\widehat{\mathbb{R}^n}$ is either a sphere in \mathbb{R}^n or the union of a plane in \mathbb{R}^n with $\{\infty\}$.

Then we can simply say:

Theorem 6.1. *Sphere inversions preserve angles and map hyperspheres in $\widehat{\mathbb{R}^n}$ to hyperspheres in $\widehat{\mathbb{R}^n}$.*

Since circles and, more generally, k -dimensional spheres for $1 \leq k < n$ are intersections of $n - k$ hyperspheres, sphere inversions preserve spheres of any dimension:

Corollary 6.2. *Sphere inversions map k -spheres in $\widehat{\mathbb{R}^n}$ to k -spheres in $\widehat{\mathbb{R}^n}$.*

Just as hyperplanes are limiting cases of hyperspheres, reflections in hyperplanes are limiting cases of sphere inversions. The reflection in the hyperplane with equation $\langle x - a, v \rangle = 0$ is the map

$$x \mapsto x' = x - 2 \frac{\langle x - a, v \rangle}{\langle v, v \rangle} v,$$

which we extend from \mathbb{R}^n to $\widehat{\mathbb{R}^n}$ by declaring that reflections in hyperplanes map ∞ to ∞ .

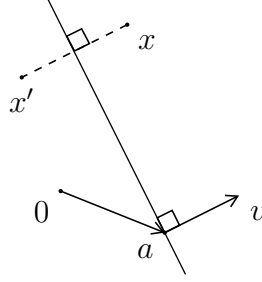


Figure 19. Reflection in a hyperplane

Definition 6.3. A Möbius transformation of $\mathbb{R}^n \cup \{\infty\}$ is a composition of sphere inversions and reflections in hyperplanes. The Möbius transformations form a group called the *Möbius group* and denoted by $\text{Möb}(n)$.

Remark 6.1. A Möbius transformation is orientation reversing or preserving depending on whether it is the composition of an odd or even number of reflections. The subgroup of orientation preserving Möbius transformations is called the *special Möbius group* and denoted by $\text{SMöb}(n)$.

Because reflections preserve angles and map spheres to spheres, Theorem 6.1 extends to Möbius transformations:

Theorem 6.3. *Möbius transformations preserve angles and map spheres in $\widehat{\mathbb{R}^n}$ to spheres in $\widehat{\mathbb{R}^n}$.*

Similarity transformations on \mathbb{R}^n are the transformations of the form $x \mapsto \lambda Ax + b$ with $\lambda > 0$, $A \in O(n)$, and $b \in \mathbb{R}^n$. Reflections in hyperplanes are a special case, and like reflections in hyperplanes we extend all similarity transformations from \mathbb{R}^n to $\widehat{\mathbb{R}^n}$ by declaring that ∞ maps to ∞ .

Proposition 6.4. *The Möbius group contains all similarity transformations.*

Proof. The group of similarity transformations is generated by translations, orthogonal transformations, and scalings.

- A translation $x \mapsto x + v$ is the composition of two reflections in parallel hyperplanes.
- An orthogonal transformation $x \mapsto Ax$ with $A \in O(n)$ is the composition of at most n reflections in hyperplanes through the origin.
- A scaling transformation $x \mapsto \lambda x$ with $\lambda > 0$ is the composition of a reflection in the unit sphere followed by a reflection in a sphere with center 0 and radius $\sqrt{\lambda}$. \square

Conversely, one only needs to add one sphere inversion to the group of similarity transformations to generate the Möbius group:

Proposition 6.5. *Every Möbius transformation is a composition of similarity transformations and inversions in the unit sphere.*

By Theorem 6.3, Möbius transformations map hyperspheres to hyperspheres. This property already characterizes all Möbius transformations.

Theorem 6.6 (Fundamental theorem of Möbius geometry). *Any bijective map $f : \widehat{\mathbb{R}^n} \rightarrow \widehat{\mathbb{R}^n}$ which maps hyperspheres in $\widehat{\mathbb{R}^n}$ to hyperspheres in $\widehat{\mathbb{R}^n}$ is a Möbius transformation.*

6.2 Two-dimensional Möbius geometry

This case is special because we can identify \mathbb{R}^2 with \mathbb{C} and $\widehat{\mathbb{R}^2}$ with the extended complex plane $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, which is the same as \mathbb{CP}^1 , the complex projective line.

- The orientation preserving and reversing similarity transformations are $z \mapsto az + b$ and $z \mapsto a\bar{z} + b$ with $a \in \mathbb{C} \setminus \{0\}$, respectively.
- Complex conjugation $z \mapsto \bar{z}$ is the reflection in the real line,
- and inversion in the unit circle $|z| = 1$ is the map $z \mapsto \frac{z}{|z|^2} = \frac{1}{\bar{z}}$.

Proposition 6.7. *The orientation preserving and reversing Möbius transformations of $\widehat{\mathbb{C}} = \mathbb{CP}^1$ are precisely the maps of the forms*

$$z \mapsto \frac{az + b}{cz + d} \quad \text{or} \quad z \mapsto \frac{a\bar{z} + b}{c\bar{z} + d} \quad \text{with} \quad ad - bc \neq 0.$$

Corollary 6.8. *The group of orientation preserving Möbius transformations of $\widehat{\mathbb{C}}$ is*

$$SMöb(2) = \text{PGL}(2, \mathbb{C}) = \text{PSL}(2, \mathbb{C}),$$

so oriented two-dimensional Möbius geometry is the same as one-dimensional complex projective geometry.

Remark 6.2. For later reference, we state the form of two important subgroups. The Möbius transformations mapping the upper half-plane to the upper half-plane are given by

$$z \mapsto \frac{az + b}{cz + d} \quad \text{or} \quad z \mapsto \frac{-a\bar{z} + b}{-c\bar{z} + d} \quad \text{with} \quad a, b, c, d \in \mathbb{R}, \quad ad - bc = 1.$$

The Möbius transformations mapping the upper half-plane to the upper half-plane are given by

$$z \mapsto \frac{az + b}{bz + \bar{a}} \quad \text{or} \quad z \mapsto \frac{a\bar{z} + b}{b\bar{z} + \bar{a}} \quad \text{with} \quad a, b \in \mathbb{C}, \quad |a| - |b| = 1.$$

The connection between two-dimensional Möbius geometry and one-dimensional complex projective geometry has the following immediate consequences:

Corollary 6.9.

- (i) *Orientation preserving Möbius transformations of $\widehat{\mathbb{C}}$ preserve the complex cross-ratio of four points*

$$\text{cr}(f(z_1), f(z_2), f(z_3), f(z_4)) = \text{cr}(z_1, z_2, z_3, z_4) = \frac{(z_1 - z_2)(z_3 - z_4)}{(z_2 - z_3)(z_4 - z_1)},$$

while orientation reversing Möbius transformations of $\widehat{\mathbb{C}}$ satisfy

$$\text{cr}(f(z_1), f(z_2), f(z_3), f(z_4)) = \overline{\text{cr}(z_1, z_2, z_3, z_4)}.$$

(ii) For any three points $z_1, z_2, z_3 \in \hat{\mathbb{C}}$ and any three points $w_1, w_2, w_3 \in \hat{\mathbb{C}}$, there is a unique orientation preserving Möbius transformation f with $f(z_i) = w_i$.

There is also a unique orientation reversing Möbius transformation g with $g(z_i) = w_i$, which is given by the composition of f followed by an inversion in the circle through w_1, w_2, w_3 , or, which is the same, inversion in the circle through z_1, z_2, z_3 followed by f .

Furthermore, the complex cross-ratio yields a convenient way to determine whether four points in the plane lie on a circle.

Proposition 6.10. *Four points z_1, z_2, z_3, z_4 lie on a circle in $\hat{\mathbb{C}}$ if and only if their cross ratio is real. Moreover, they lie on a circle in that cyclic order if and only if $\text{cr}(z_1, z_2, z_3, z_4) < 0$.*

The fixed points of an orientation preserving Möbius transformation f are obtained by solving

$$f(z) = \frac{az + b}{cz + d} = z,$$

which (if $c \neq 0$) is a quadratic equation in z given by

$$cz^2 - (a - d)z - b = 0.$$

It has at least one and at most two (if $f \neq \text{id}$), given by

$$z_{\pm} = \frac{(a - d) \pm \sqrt{\Delta}}{2c}$$

with discriminant

$$\Delta = (a - d)^2 + 4bc = (a + b)^2 - 4(ad - bc) = (\text{tr } F)^2 - 4 \det F,$$

where

$$F = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \text{GL}(2, \mathbb{C})$$

is a matrix representation of f . If we choose a representation $F \in \text{SL}(2\mathbb{C})$, i.e., $\det F = 1$, the trace of F is uniquely determined up to sign. We obtain

$$\begin{aligned} f \text{ parabolic} & \quad :\Leftrightarrow \quad f \text{ has exactly one fixed point} \\ & \quad \Leftrightarrow \quad (\text{tr } F)^2 = 4, \end{aligned}$$

and

$$\begin{aligned} f \text{ non-parabolic} & \quad :\Leftrightarrow \quad f \text{ has two distinct fixed point} \\ & \quad \Leftrightarrow \quad (\text{tr } F)^2 \neq 4. \end{aligned}$$

In both cases, we may bring f to one of the following normal forms.

Proposition 6.11. *Let $f \neq \text{id}$ be an orientation preserving Möbius transformation.*

(i) *If f is non-parabolic, then there exists a Möbius transformation g and $k \in \mathbb{C} \setminus \{0, 1\}$ such that*

$$(g \circ f \circ g^{-1})(z) = kz.$$

(ii) If f is parabolic, then there exists a Möbius transformation g and $\beta \in \mathbb{C} \setminus \{0\}$ such that

$$(g \circ f \circ g^{-1})(z) = z + \beta.$$

Proof.

(i) Let $\gamma_1, \gamma_2 \in \mathbb{C}$ be the fixed points of f . Choose

$$g(z) = \frac{z - \gamma_1}{z - \gamma_2},$$

which satisfies $g(\gamma_1) = 0$ and $g(\gamma_2) = \infty$.

(ii) Let $\gamma \in \mathbb{C}$ be the fixed point of f . Choose

$$g(z) = \frac{1}{z - \gamma},$$

which satisfies $g(\gamma) = \infty$.

□

For non-parabolic Möbius transformation f the number k is called the *multiplier* of γ_1 . An $\text{SL}(2, \mathbb{C})$ representative of the normal form is given by

$$\tilde{F} = \begin{pmatrix} \lambda & 0 \\ 0 & \frac{1}{\lambda} \end{pmatrix}, \quad k = \lambda^2.$$

Thus, we obtain the following relation between the multiplier k and the trace

$$(\text{tr } F)^2 = \left(\lambda + \frac{1}{\lambda} \right)^2 = \left(\sqrt{k} + \frac{1}{\sqrt{k}} \right)^2$$

The multiplier is invariant under conjugation, and thus

$$f'(\gamma_1) = k.$$

Interchanging the two fixed points $\gamma_1 \leftrightarrow \gamma_2$ leads to $k \leftrightarrow \frac{1}{k}$, and thus

$$f'(\gamma_2) = \frac{1}{k}.$$

In particular, if one fixed point is attractive, then the other fixed point is repelling. One defines

$$\begin{aligned} f \text{ elliptic} & : \Leftrightarrow |k| = 1 & \Leftrightarrow (\text{tr } F)^2 \in [0, 4), \\ f \text{ loxodromic} & : \Leftrightarrow |k| \neq 1 & \Leftrightarrow (\text{tr } F)^2 \notin [0, 4], \\ f \text{ hyperbolic} & : \Leftrightarrow k \in \mathbb{R} & \Leftrightarrow (\text{tr } F)^2 \in (4, \infty). \end{aligned}$$

From this definition and Proposition 6.11, we can easily conclude the following decomposition of orientation preserving Möbius transformations into inversions in circles.

Proposition 6.12.

(i) Every loxodromic Möbius transformation is the composition of an elliptic and a hyperbolic Möbius transformation.

- (ii) Every parabolic, elliptic, and hyperbolic Möbius transformation is the composition of two inversions (or reflections).
- (iii) Every orientation preserving Möbius transformation is the composition of at most four inversions (or reflections).

An non-parabolic Möbius transformation is uniquely determined by its two fixed points γ_1, γ_2 and the multiplier k .

Proposition 6.13. *Let $f \neq \text{id}$ be an orientation preserving Möbius transformation.*

- (i) *If f is non-parabolic, and γ_1, γ_2 its two fixed points, and k the multiplier (of γ_1), then*

$$F(k; \gamma_1, \gamma_2) = \begin{pmatrix} \gamma_1 - k\gamma_2 & (k-1)\gamma_1\gamma_2 \\ 1-k & k\gamma_1 - \gamma_2 \end{pmatrix} \in \text{SL}(2, \mathbb{C}).$$

In the case $\gamma_2 = \infty$

$$F(k; \gamma_1, \infty) = \begin{pmatrix} k & (1-k)\gamma_1 \\ 0 & 1 \end{pmatrix} \in \text{SL}(2, \mathbb{C}).$$

- (ii) *If f is parabolic, and $\gamma \in \mathbb{C}$ its fixed point, and β the translation length, then*

$$F(\beta; \gamma) = \begin{pmatrix} 1 + \gamma\beta & -\beta\gamma^2 \\ \beta & 1 - \gamma\beta \end{pmatrix} \in \text{SL}(2, \mathbb{C}).$$

In the case $\gamma = \infty$

$$F(\beta; \infty) = \begin{pmatrix} 1 & \beta \\ 0 & 1 \end{pmatrix} \in \text{SL}(2, \mathbb{C}).$$

Proof.

- (i) Using the function g from the proof of Proposition 6.11, we have

$$g \circ f(z) = kg(z) \quad \Leftrightarrow \quad \frac{f(z) - \gamma_1}{f(z) - \gamma_2} = k \frac{z - \gamma_1}{z - \gamma_2}$$

Solving for $f(z)$ yields the result.

- (ii) Similarly,

$$g \circ f(z) = g(z) + \beta \quad \Leftrightarrow \quad \frac{1}{f(z) - \gamma} = \frac{1}{z - \gamma} + \beta.$$

□

For a non-parabolic Möbius transformation f with fixed points $\gamma_1, \gamma_2 \in \mathbb{C}$, let us note the elliptic pencil of circles through γ_1 and γ_2 by

$$\mathcal{C}_e(\gamma_1, \gamma_2) := \{\text{circles containing } \gamma_1 \text{ and } \gamma_2\},$$

and the hyperbolic pencil of orthogonal circles by

$$\mathcal{C}_h(\gamma_1, \gamma_2) := \{\text{circles orthogonal to all circles of } \mathcal{C}_h(\gamma_1, \gamma_2)\}.$$

Then

- ▶ An elliptic transformation will map each circle from \mathcal{C}_e to another circle from \mathcal{C}_e , while preserving each circle in \mathcal{C}_h . ^{JT:} [how does the angle relate to k ?]
- ▶ A hyperbolic transformation will map each circle from \mathcal{C}_h to another circle from \mathcal{C}_h , while preserving each circle in \mathcal{C}_e . ^{JT:} [relation inversive distance and k]
- ▶ A loxodromic transformation will map each circle from \mathcal{C}_h to another circle from \mathcal{C}_h , but not preserve the circles in \mathcal{C}_e . Instead it preserves loxodromic curves connecting γ_1 , and γ_2 . These are curves of constant angle with the circles from \mathcal{C}_h , or equivalently, Möbius images of logarithmic spirals. ^{JT:} [should say much more about these curves...] ^{JT:} [what is this angle, should be again be related to k]

We may say a little more about this pairing of circles in $\mathcal{C}_h(\gamma_1, \gamma_2)$ by a loxodromic transformations. To this end, let us distinguish the inside and outside of the circles (Euclidean distinction). The family $\mathcal{C}_h(\gamma_1, \gamma_2)$ may be separated into two components by the perpendicular bisector of γ_1 and γ_2 .

- ▶ If f maps a circle C_1 of one component to a circle C_2 of the same component, it maps the inside of C_1 to the inside of C_2 .
- ▶ If f maps a circle C_1 of one component to a circle C_2 of the other component, it maps the inside of C_1 to the outside of C_2 .

In the second case the two circles C_1 and C_2 are called *paired* by f .

Definition 6.4. Two circles $C_1, C_2 \subset \mathbb{C}$ non contained inside the other are called (*Schottky-paired*) by the orientation preserving Möbius transformation f if f maps the inside of C_1 to the outside of C_2 (and thus C_1 to C_2 and the outside of C_1 to the inside of C_2).

Remark 6.3. If we cut out the inside of C_1 and C_2 and identify points of C_1 and C_2 that are mapped to each other by f , the resulting surface is a topological torus.

Using the formula for a Möbius transformation from fixed points and multiplier (Proposition 6.13), which circle in $\mathcal{C}_h(\gamma_1, \gamma_2)$ is mapped to which is determined by the absolute value $|k|$ of the multiplier. ^{JT:} [however, this still doesn't give the general pairing.]

However, given two circles C_1 and C_2 non contained inside the other, how can we Schottky-pair them by a Möbius transformation? We can first translate and scale C_1 to the unit circle, then invert in the unit circle, and then scale and translate the unit circle to C_2 . By inserting a general Möbius transformation in between, that maps the unit disk to itself, we obtain the most general form of such a transformation.

Proposition 6.14. Let $C_1, C_2 \subset \mathbb{C}$ be two circles with centers, $c_1, c_2 \in \mathbb{C}$ and radii $r_1, r_2 > 0$. Then a general orientation preserving Möbius transformations, that maps C_1 to C_2 is given by

$$z \mapsto r_2 \frac{\bar{b}(z - c_1) + r_1 \bar{a}}{a(z - c_1) + r_1 b} + c_2 \quad \text{with } a, b \in \mathbb{C}, |a| - |b| = 1.$$

^{JT:} [what does all of this look like in the projective model?]

6.3 Schottky groups and limit sets

Definition 6.5. A discrete subgroup (no limit points) of $\mathrm{PSL}(2, \mathbb{C})$ is called a *Kleinian group*.

Remark 6.4. A discrete subgroup of $\mathrm{PSL}(2, \mathbb{R})$ is called a *Fuchsian group*. Thus, every Kleinian group that preserves the real line is a Fuchsian group, and every Kleinian group that preserves a circle is conjugate to a Fuchsian group.

Definition 6.6. Let $C_1, \tilde{C}_1, \dots, C_g, \tilde{C}_g \subset \mathbb{C}$ be $2g$ circles with disjoint interiors. and $f_1, \dots, f_g \in \mathrm{PSL}(2, \mathbb{C})$ Möbius transformations such that C_i and \tilde{C}_i are (Schottky-)paired by f_i for $i = 1, \dots, g$, then the Kleinian group generated by f_1, \dots, f_g is called a (*classical*) *Schottky group*.

Remark 6.5. A fundamental domain F for the action of a Schottky group G on $\hat{\mathbb{C}}$ is given by the common exterior of all the circles $C_1, \tilde{C}_1, \dots, C_g, \tilde{C}_g$. The quotient F/G is a compact Riemann surface of genus g .

For simplicity, from now on we consider Schottky groups generated by two loxodromic Möbius transformations. We introduce the following notations, and make some observations, following [Indra's pearls - David Mumford, Caroline Series, David Wright]:

- We denote the two generators of the group by

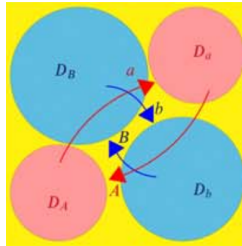
$$a, \quad b,$$

and its inverse transformations by

$$A := a^{-1}, \quad B := b^{-1}.$$

- We denote the circles paired by a by C_A and C_a , and their interior disks by D_A and D_a . Thus, a maps

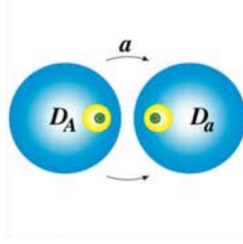
$$\begin{aligned} a(C_A) &= C_a, & a(\hat{\mathbb{C}} \setminus \overline{D_A}) &= D_a, & a(D_A) &= \hat{\mathbb{C}} \setminus \overline{D_a}, \\ A(C_a) &= C_A, & A(\hat{\mathbb{C}} \setminus \overline{D_a}) &= D_A, & A(D_a) &= \hat{\mathbb{C}} \setminus \overline{D_A}. \end{aligned}$$



Successive application of a or A leads to nested disks, which we call

$$\begin{aligned} D_{\underbrace{a \cdots a}_{k+1}} &:= \underbrace{a \cdots a}_k(D_a) \subset D_{\underbrace{a \cdots a}_k} \subset \cdots \subset D_{aa} \subset D_a, \\ D_{\underbrace{A \cdots A}_{k+1}} &:= \underbrace{A \cdots A}_k(D_A) \subset D_{\underbrace{A \cdots A}_k} \subset \cdots \subset D_{AA} \subset D_A. \end{aligned}$$

These two sequences converge to the attractive fixed point of a and A respectively.



Similarly, for b, B, C_B, C_b, D_B, D_b .

- Every element of the Schottky group is represented by a sequence

$$\sigma_1 \cdots \sigma_k$$

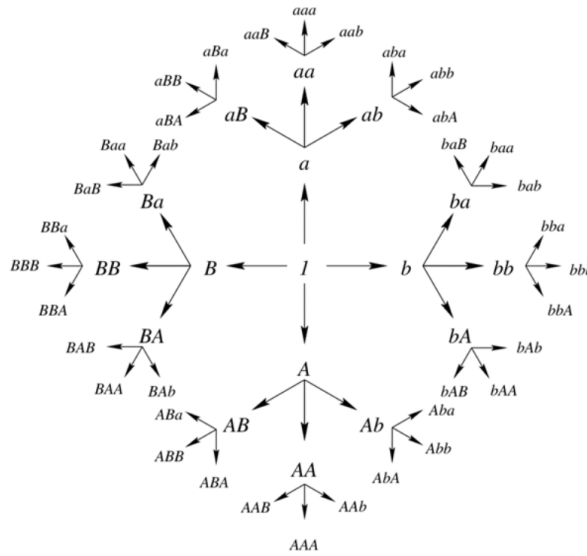
sometimes called a *word* consisting of the letters σ_i given by

$$a, A, b, B.$$

The only relations in the group satisfied are

$$aA = Aa = bB = Bb = 1,$$

which lead to non-unique representations and possible cancellations of letters in a word. If all possible cancellations are applied (no adjacent a, A and no adjacent b, B), the word is called *reduced*.



All reduced words are represented by this *word tree*.

- We now apply all elements of the Schottky group to the initial disks D_a, D_A, D_b, D_B , and denote the images by

$$D_{\sigma_1 \cdots \sigma_{k+1}} := \sigma_1 \cdots \sigma_k(D_{\sigma_{k+1}}) \subset D_{\sigma_1 \cdots \sigma_k} \subset \cdots \subset D_{\sigma_1 \sigma_2} \subset D_{\sigma_1}.$$

For example applying the transformation a to the disks D_a, D_b, D_B leads to 3 new disks

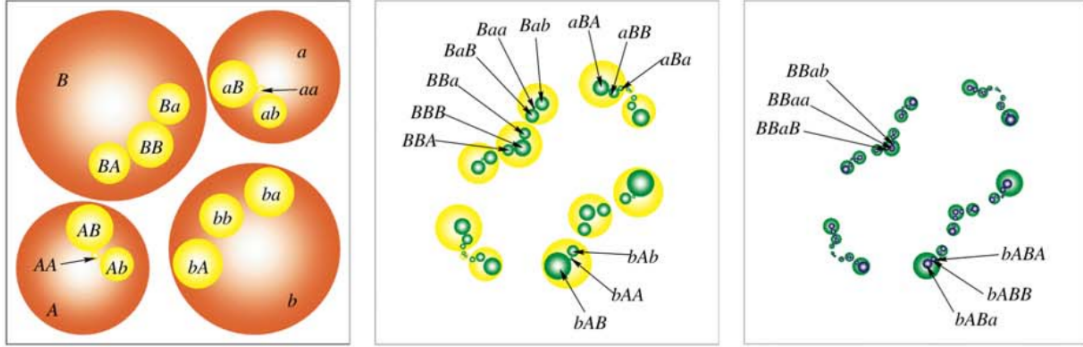
$$D_{aa}, D_{ab}, D_{aB} \subset D_a$$

all contained in D_a . Similarly,

$$\begin{aligned} D_{AA}, D_{Ab}, D_{AB} &\subset D_A, \\ D_{ba}, D_{bA}, D_{bb} &\subset D_b, \\ D_{Ba}, D_{BA}, D_{BB} &\subset D_B, \end{aligned}$$

which brings us to 12 disks on the second level.

9 of these 12 disks lie outside D_A , thus applying a to those disks, yields 9 disks contained in D_a , 3 contained in D_{aa} , 3 in contained in D_{ab} , and 3 contained in D_{aB} . In total, we obtain 36 disks on the third level.



- The collection of all disks obtained in that way

$$\{D_{\sigma_1 \dots \sigma_k} \mid \sigma_1 \dots \sigma_k \text{ word}\}$$

is a “pattern” or “tiling” called *Schottky array*. ^{JT:} [instead of using words, here we can just say its obtained by applying any element of the group to an initial disk] It is invariant under applying the Schottky group.

In particular, we call the collection of disks of the k -th level

$$\mathcal{D}_k := \bigcup_{\sigma_1 \dots \sigma_k \text{ reduced word}} D_{\sigma_1 \dots \sigma_k}$$

the *level- k Schottky array*.

- The set of points belonging to a disk of every level of the Schottky array

$$F = \bigcap_{k=1}^{\infty} \mathcal{D}_k$$

is called them *limit set* of the Schottky group. It is again invariant under the Schottky group.

- The level- k Schottky arrays are a decreasing sequence of sets, leading to a Cantor set-like construction. In fact the radii are decreasing exponentially fast, approaching points. As long as the initial disks do not touch, this leads to a totally disconnected set, which has Hausdorff dimension $0 < \dim_H F < 1$.

- Each such *limit point* corresponds to an infinite (reduced) word, coming from a nested sequence of disks

$$D_{\sigma_1\sigma_2\sigma_3\cdots} \subset \cdots \subset D_{\sigma_1\sigma_2\sigma_3} \subset D_{\sigma_1\sigma_2} \subset D_{\sigma_1}$$

In particular, every periodic word

$$\bar{w} = www\cdots, \quad w = \sigma_1 \cdots \sigma_p$$

corresponds to a nested sequence

$$D_{\bar{w}} \subset \cdots \subset D_{www} \subset D_{ww} \subset D_w,$$

which converges to the attractive fixed point of w . Thus, the limit set contains all attractive fixed points of elements of the Schottky group.

Furthermore it contains all images of attractive fixed points under elements of the Schottky group, in particular all images of the four attractive fixed points of a, A, b, B . Those correspond to words of the form $w\bar{a}$ etc.

- This leads to the following ways of visualizing the limit set of a Schottky group: Let $N \in \mathbb{N}$ large.
 - Plot the disks D_w for all reduced words w of length N .
 - For some point $z_0 \in \mathbb{C}$ plot $w(z_0)$ for all reduced words w of length N .
 - Plot all fixed points of words of length at most N .
 - Plot all points $w(\gamma_i)$ for all reduced words w of length at most N , where γ_i are the attractive fixed points of $i = a, A, b, B$. (Alternatively choose another finite set of attractive fixed points of elements of the Schottky group.)

- Special case: All initial disks orthogonal to a common circle.

In this case all disks of the Schottky array are orthogonal to this circle, and thus the limit set is contained in the circle.

The Schottky group is (conjugate to) a Fuchsian group in this case.

- Special case: The initial disks D_a, D_b, D_A, D_B touch cyclically, and the generators match the touching points

In this case each level- k Schottky array is a chain of touching disks, and the limit set becomes a (fractal) curve.