Geometry

by

V. I. Arnold ,
R. Uribe-Vargas

ΑΓΕΩΜΕΤΡΗΤΟΣ ΜΗΔΕΙΣ ΕΙΣΙΤΩ *

No one unversed in geometry should enter here

Preface

The French Encyclopedic Dictionary « Le Petit Larousse » (1994) defines Mathematics as: « Science which studies the properties of abstract entities, and the relations between them, by means of deductive reasonings ».

Thus, no relation of mathematics to the real world is mentioned.

In the present book, Mathematics is presented as a (natural) science which studies the most fundamental properties of the objects of the real world (perceived experimentally), by means of inductive reasonings.

The shortest way to understand what is this book about is to contemplate its multiple pictures. We present geometry as the part of mathematics dealing with the visual images (including the "geometry of numbers" and "geometry of formulae" in number theory and algebra).

Geometry is the science that studies the real figures and shapes rather than the sequences of abstract symbols. It provides a mathematical (that is, a precise) objective description of the real world, discerning those properties of the observed shapes which are independent of the illusions of the particular observer. All applications of mathematics are based on this geometric understanding of the reality.

Rousseau claimed that he understood the formula

$$(a+b)^2 = a^2 + 2ab + b^2$$

not at all when he obtained it by the multiplication of two brackets, but rather when the drawing of rectangular areas (see Fig. 1) persuaded him that it is natural. Fig. 1 shows also two other geometric reasonings: Pythagoras theorem (known and written with the proof several thousand years before him) and the evaluation of the area of the circular disc $S = \pi r^2$.

*The legendary inscription at the entrance to the "Academy" of Plato, which existed 918 years (between -385 and $+529$).

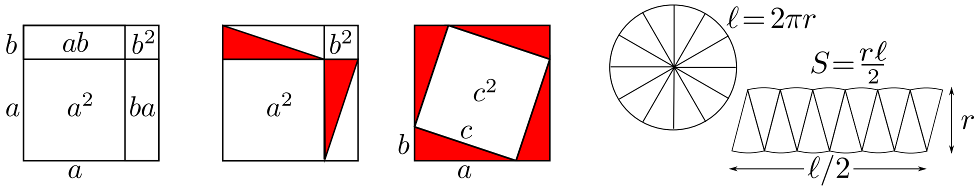


Figure 1: Three celebrated geometrical theorems.

The authors of the present book were trying to make many other geometric facts as clear as these celebrated theorems, like the statement of Fig. 2:

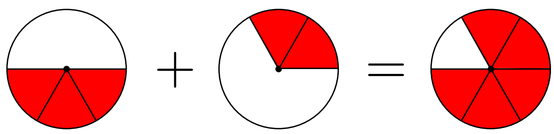


Figure 2: Fractions addition theory (Poincaré's apple).

$$\frac{1}{2} + \frac{1}{3} = \frac{5}{6}$$

(contradicting the usual opinion of the people inclined to algebra that this sum is $\frac{2}{5}$, since $1 + 1 = 2$ and $2 + 3 = 5$).

We shall try to make for the reader as clear as this elementary theory the fact that a hedgehog cannot be brushed, due to its spherical form (while it would be possible if the hedgehog surface were toric – Fig. 3).

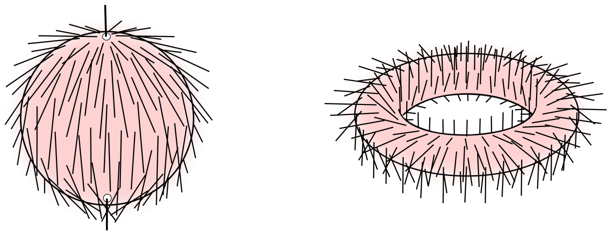


Figure 3: Two undressed needles of a spherical hedgehog and a dressed torical hedgehog.

This result belongs to the topological part of geometry. The book explains the main notions of topology, like homotopy groups and homology rings, the theories of knots and braids, the linking numbers and the exact sequences, the spectral sequences and the vector bundles.

Together with the geometric theory of the Riemannian tensor of curvature we shall describe the symmetry groups of the platonic regular bodies (used by Kepler to describe the axes of the Solar system orbits), the relativistic meaning of the notion of normal subgroups of algebra, the relation of Lobachevsky geometry to the de Sitter relativistic worlds and to the Poisson brackets in the Lie algebra of the quadratic forms on a symplectic plane, the helicity and Hopf invariants of magnetic fields and the Lie "fisherman" derivatives of the differential forms, the elliptic coordinates of Jacobi and the topological proof of Abel's theorem on the unsolvability of the algebraic equations of degree 5 by radicals. The reader will understand the Milnor's exotic spheres, the representation of the Euler characteristic of the surfaces in terms of the integral of their Gaussian curvatures, and the quaternion and spin geometries.

Modern geometry of manifolds has became the base of a fast development of several domains of physics, where Morse theory (of the instantons connecting the critical points) and the theory of the characteristic classes of the fibrations are today as commonly used in modern quantum fields theory, as the differentiation and the integration of the calculus.

As Jacobi stated, one of the wonderful miracles of mathematics is the universality of its results, where «the same function describes the lengths of the ellipsis, the number of representations of an integer as the sum of four squares and the genuine motion of a pendulum». We hope the reader will find in our book many examples of this universal applicability of Geometry.

The best way to understand geometry is to solve independently a hundred of its problems, of which we present many lists in Chapter 19.

The written examinations experience shows that these problems are accessible for the new-coming students, while most professors, spoiled by the axiomatico-deductive educational methods, are unable to solve them. These problems need rather the natural science approach, where the examples teach more than the rules and than the formal definitions*.

The reader of the present book is not expected to be a mathematician, but he is expected to understand some kinder-garden mathematics and to

---

*Some modern professors claim that "geometry is the part of algebra distinguished by the convention not to write the summation symbol" mentioning the Einstein's notations. Hilbert claimed in 1930 that the geometry is a part of physics, leading some mathematicians to conclude later that it has no relation to mathematics.

see mathematics as a real world study, rather than as a game with formal sequences of symbols. Thus we suppose that the reader understand that $2 + 3 = 5$ (rather than that $2 + 3 = 3 + 2$, which he studied at school) and that the number $4/7$ is smaller than 1. The reader should understand that a cube has 8 vertices and 6 faces, that a vector is an arrow, the determinant - the volume, the derivative - the velocity, the integral - the area.

We shall use the number π (defined in Fig. 1) and the Euler number e, the functions sin and cos, the exponential and the logarithms, the eigenvalues of matrices and the quadratic forms, the ellipses and the hyperbolas.

Unfortunately, in the modern educational systems many basic notions and ideas are hidden from the students in the abstract axioms (to enhance the authority of the teachers and of their teaching).

To help such students we shall explain with many details some basic elements of mathematics (see Fig. 2 above), providing several examples and problems along the book.

This approach makes that many domains of mathematics, which might look difficult and mysterious in the dogmatic expositions, become transparent and natural, leaving their inaccessibility halo.

Contents

Chapter 1

Manifolds and maps of manifolds

Even in the mathematical sciences,
our principal instruments to discover the truth
are induction and analogy.

Pierre-Simon de Laplace

1.1 Manifolds

The main objects in geometry are the manifolds. A circle, a straight line, a parabola and a hyperbola in the plane, given by the respective equations $x^2 + y^2 = 1$, $x + y = 1$, $y = x^2$ and $x^2 - y^2 = 1$, are examples of manifolds.

Definition. A subset M^k of the space $\mathbb{R}^n$ (or $\mathbb{C}^n$) is a *submanifold* of dimension k if for every point $p \in M^k$ there exist a decomposition $\mathbb{R}^n = \mathbb{R}^k \oplus \mathbb{R}^{n-k}$, a map $f_p : \mathbb{R}^k \to \mathbb{R}^{n-k}$ and a neighbourhood U of p such that

$$M^k \cap U = \{(x, y) \in U : y = f_p(x)\} ,$$

where x and y denote the respective projections on $\mathbb{R}^k$ and $\mathbb{R}^{n-k}$. Such a decomposition depends on the point p.

If all the functions f_p covering M are of class C^r, the submanifold is said to be *of class* C^r.

Since the submanifold M is locally the graph of a function f_p, near the point p it is parametrised by the map $x \mapsto (x, f_p(x))$. The pair formed by the domain $\overline{U}$ of $\mathbb{R}^k$ over which f_p is defined together with the map $\overline{U} \to \mathbb{R}^n$, $x \mapsto (x, f_p(x))$ is called a *local chart* of M.

Example. The *sphere* of dimension n, $\mathbb{S}^n \subset \mathbb{R}^{n+1}$, defined by the equation $x_1^2 + \cdots + x_{n+1}^2 = 1$, is an important example of smooth submanifold. Notice that it is not possible to cover the sphere with only one global chart. In the simplest case of the circle ($n = 1$) one needs at least three such charts – Fig. 1.1. Near the point p_1 of coordinates (x_1, y_1), the circle can be viewed as the graph $y = f(x)$ of the smooth function $f(x) = \sqrt{1 - x^2}$, but also as the graph $x = g(y)$ of the smooth function $g(y) = \sqrt{1 - y^2}$. Near the second point $p_2 = (1, 0)$, the circle can be viewed as the graph of g, but not of f.

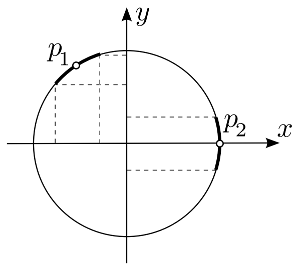


Figure 1.1: Local charts of the circle.

When a point in a manifold is covered by two charts, there exist smooth maps transforming the coordinates in a chart to those in the other chart. These maps are called *coordinate changes*.

A set of charts covering a manifold, together with the corresponding coordinate changes, is called an *atlas* of that manifold.

Smooth maps. Let us consider two smooth submanifolds M and N of dimensions m and n, and a map $f : M \to N$. In local charts, f is represented by some map from some domain of $\mathbb{R}^m$ to a domain of $\mathbb{R}^n$. The map f is smooth if all its representations in these local charts are smooth maps.

Smoothness. Here and in the sequel unless there is explicit mention to the contrary the word *smooth* or *differentiable* means "continuously differentiable the necessary number of times", for example infinitely differentiable.

Diffeomorphism. Suppose the two submanifolds have the same dimension and $f : M \to N$ has an inverse map $f^{-1} : N \to M$. If both f and f^{-1} are smooth (or continuous), then f is a *diffeomorphism* (resp. *homeomorphism*).

Examples. **1**. The C^∞ map $g : x \mapsto \ln x$ is a diffeomorphism of the half line (of positive real numbers) onto the whole line $\mathbb{R}$. Its inverse is $y \mapsto e^y$.

2. The function $f : x \mapsto x^3$ is of class C^∞ but its inverse $y \mapsto y^{1/3}$ is only continuous. Thus, f is a homeomorphism but it is not a diffeomorphism.
3. The map $f : (x, y) \mapsto (X, Y) = (x, x^2 - y)$ is a diffeomorphism of the plane $\mathbb{R}^2$ onto itself. Indeed f is its own inverse: $f^{-1} = f$ (verify it!).

Diffeomorphic submanifolds. Two given submanifolds are said to be *diffeomorphic* (resp. *homeomorphic*) if there exists a global diffeomorphism (resp. homeomorphism) from one to the other.

Thus two homeomorphic submanifolds may be non diffeomorphic.

Examples. **1**. The circle $x^2 + y^2 = 1$ and the ellipse $(X/a)^2 + (Y/b)^2 = 1$ are diffeomorphic under $(x, y) \mapsto (X, Y) = (ax, by)$.
2. In Euclidean plane the circle and the hyperbola $x^2 - y^2 = 1$ are not homeomorphic and hence they are not diffeomorphic.
3. In Euclidean plane the line $y = 0$ and the parabola $Y = X^2$ are diffeomorphic under the map $(x, y) \mapsto (X, Y) = (x, x^2 - y)$.
4. A circle is homeomorphic but not diffeomorphic to a square. That is, there exist no homeomorphism of a circle onto a square, such that the functions defining it locally are diffeomorphisms from an arc of circle onto a part of the square, for a neighbourhood of every point.
5. The sphere $\mathbb{S}^2$ and the torus $T^2 := \mathbb{S}^1 \times \mathbb{S}^1$ are neither homeomorphic nor diffeomorphic (since they have different numbers of "holes").

Definition. The *abstract manifolds* are the equivalence classes of the submanifolds under diffeomorphisms.

For example, a circle and an ellipse represent the same abstract manifold.

Hence, we say that an abstract manifold M is *embedded* in Euclidean space $\mathbb{R}^N$ if we are taking a submanifold of $\mathbb{R}^N$, which is a representative of the equivalence class of M.

1.2 Classification of Closed Surfaces

An important problem in geometry and topology is the classification of (abstract) manifolds. For 1-dimensional manifolds (that is, for *curves*), the classification is given by the following theorem.

Theorem. *The only connected closed smooth curve in $\mathbb{R}^n$ is the circle.*

One often encounters the so-called *closed manifolds*: compact and without boundary. The sphere $\mathbb{S}^2$ and the torus $\mathbb{T}^2$ are examples of closed manifolds.

From the topological viewpoint, the torus is just a sphere with a handle. Similarly, we get a sequence of non diffeomorphic (and also non homeomorphic) surfaces, by attaching g handles to the sphere – Fig. 1.2.

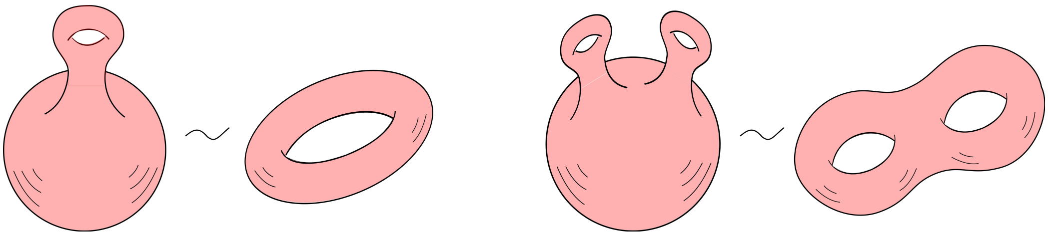


Figure 1.2: The sphere with $g = 1, 2$ handles.

Looking for the classification of connected closed surfaces (2-dimensional manifolds), one may wonder whether every connected closed smooth surface is diffeomorphic to a sphere with g handles.

The answer to this problem is given by the surface called *Möbius band*, obtained from a rectangle by gluing two sides as indicated in Fig. 1.3.

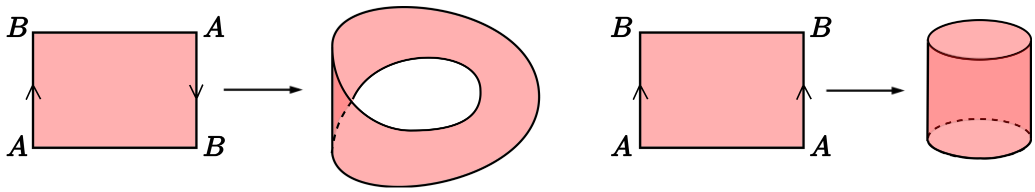


Figure 1.3: The Möbius band and the cylinder.

Strictly speaking, the Möbius band is a surface with boundary. Its boundary is a circle $\mathbb{S}^1$. However, by gluing a disc to the Möbius band along this circle, we obtain a closed surface which is not a sphere with g handles.

Remark. The other possible gluing of two opposite sides of the rectangle provides a cylinder – Fig. 1.3. Cutting a cylinder along a horizontal line as shown in Fig. 1.4, we get a disconnected surface formed by two cylinders. However, cutting in the same way a Möbius band, we get a cylinder.

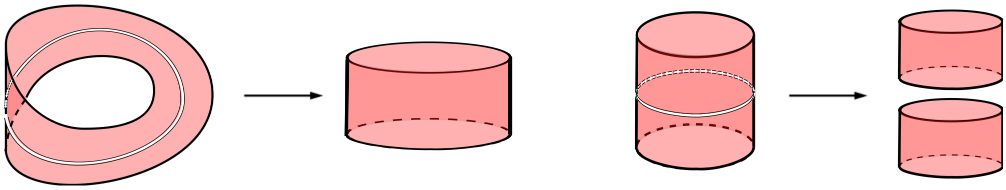


Figure 1.4: Cutting the Möbius band and the cylinder.

Orientation

The main difference between the Möbius band and the spheres with g handles is explained by the notion of orientability. In order to state an important theorem on the classification of connected closed 2-manifolds we need the

Definition. Two ordered bases of $\mathbb{R}^n$ define the *same orientation* of $\mathbb{R}^n$ if there exists a continuous path joining them in the space of the ordered bases.

Example. Consider the three ordered bases of the line $\mathbb{R}$ and of the plane $\mathbb{R}^2$, shown in Fig. 1.5. The ordered bases at A and B have the same orientation, which is different from the orientation of the ordered basis at C.

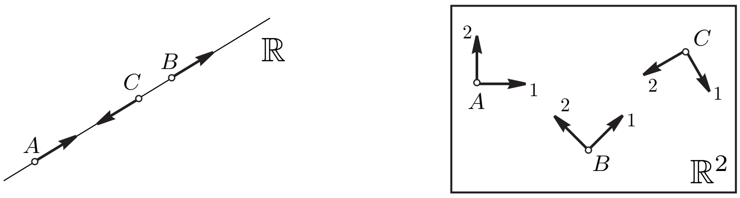


Figure 1.5: Orientations on the line and on the plane.

Theorem. *There are only two possible orientations in* $\mathbb{R}^n$.

The orientation is related to the permutations in the following way. Let σ be a permutation of n elements. The ordered bases $\{e_1, \ldots, e_n\}$ and $\{e_{\sigma(1)}, \ldots, e_{\sigma(n)}\}$ define the same orientation if and only if the permutation σ is even.

An orientation on $\mathbb{R}^n$ induces orientations on the local charts of a manifold.

Definition. A manifold is said to be *orientable* if there exists an atlas of charts such that the orientation is continuous along the manifold. A manifold is *oriented* whenever an orientation is fixed on it.

Problem. Prove that *the Möbius band is not orientable.*

Solution. Fix a coordinate system in a neighbourhood of a point of the equator of the Möbius band. Moving this coordinate system around the equator of the band, we come back to the initial point, but we obtain a coordinate system which defines the opposite orientation to that of the initial coordinate system.

Returning to the classification of connected closed surfaces, we state without proof the following difficult result due to Möbius (1870):

Theorem. *Every oriented connected closed smooth surface is diffeomorphic to a sphere with g handles.*

Genus. The number g of handles is called the *genus* of the surface. The *genus* is also the maximum number of cuttings that can be done along non-intersecting closed simple curves on the surface, keeping it connected.

Examples. **1**. The genus of the sphere is $g = 0$ because it becomes disconnected by any cutting with a circle.
2. If we cut the torus along a circle not bounding a disc, the result is a cylinder. This surface is connected but any additional cutting would disconnect it. So the genus of a torus is $g = 1$.

To give the complete list of closed connected surfaces, including the non orientable ones, we need to study the "projective plane" (obtained from the Möbius band and a disc by attaching them along their boundaries).

1.3 Projective Spaces

The projective space is the mathematical modelling of perspective.
Paolo Uccello

Definition. The n-dimensional *real projective space* $\mathbb{R}\mathrm{P}^n$ is the set whose points are the lines (or directions) of $\mathbb{R}^{n+1}$ passing through the origin:

$$\mathbb{R}\mathrm{P}^n = \frac{\mathbb{R}^{n+1} \smallsetminus \{0\}}{\mathbb{R} \smallsetminus \{0\}} \ .$$

It is interesting that Göthe wrote in a poem: *"if you would step into the infinite, you have only to walk in the finite in all directions"*.

The same way as the elements of the projective space are lines (directions), some times one has to deal with sets whose elements are not points in Euclidean space. To consider such a set as a manifold, one has to embed it as a k-dimensional submanifold in some Euclidean space $\mathbb{R}^N$. It is often useful to parametrise such spaces (at least locally) by open subsets of $\mathbb{R}^k$:

Charts and Local Coordinates. A *chart* or *local parametrisation* of a set M is an open set U in Euclidean (coordinate) space $\mathbb{R}^k$ $(x = (x_1, \ldots, x_k))$ together with a bijective map f of U onto a subset of M, $f : U \to f(U) \subset M$.

The inverse of a chart $f^{-1} : f(U) \to U \subset \mathbb{R}^k,\ \ p \mapsto (x_1(p), \ldots, x_k(p))$, is called a *coordinate system* on $f(U) \subset M$; the k functions $(x_1, \ldots, x_k)$ defined by it are called *local coordinates* in $f(U)$.

Example. The projective line $\mathbb{R}\mathrm{P}^1$ is the set of lines passing through the origin of $\mathbb{R}^2$. Equipping $\mathbb{R}^2$ with coordinates x and y, the *affine chart* $\varphi : \mathbb{R} \to \mathbb{R}\mathrm{P}^1$ that sends $\lambda \in \mathbb{R}$ to the line $\varphi(\lambda) = \ell_\lambda \in \mathbb{R}\mathrm{P}^1$ of equation $y = \lambda x$ turns the parameter λ into an *affine coordinate* of $\mathbb{R}\mathrm{P}^1$, covering all elements of $\mathbb{R}\mathrm{P}^1$ except the vertical line. As the line ℓ_λ approaches the vertical position, λ tends to ∞; so for this chart the vertical line corresponds to $\lambda = \infty$. Thus, $\mathbb{R}\mathrm{P}^1 = \mathbb{R} \sqcup \{\infty\} = \mathbb{S}^1$, where the symbol $\sqcup$ means "disjoint union".

(Observe that the intersections of the vertical line $x = 1$ with the lines passing through the origin provide the affine coordinate $\varphi^{-1} : \mathbb{R}\mathrm{P}^1 \to \mathbb{R}$ (Fig. 1.6).)

To parametrise the lines near the vertical one, $x = 0$, we take new coordinates $\widetilde{y} = x$, $\widetilde{x} = y$ and the chart ψ sending $\widetilde{\lambda} \in \mathbb{R}$ to the line $\widetilde{y} = \widetilde{\lambda}\widetilde{x}$. The two charts φ and ψ cover the whole projective line.

Another way to see that $\mathbb{R}\mathrm{P}^1 = \mathbb{S}^1$, is to parametrise the lines in $\mathbb{R}^2$ passing through the origin by the angle φ formed with the x-axis. This angle is defined modulo π. Hence, $\mathbb{R}\mathrm{P}^1 \simeq \{\varphi \pmod{\pi}\} \simeq \mathbb{S}^1$.

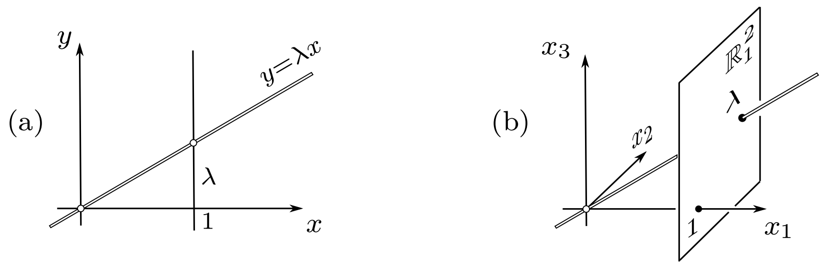


Figure 1.6: (a) Affine charts of $\mathbb{R}\mathrm{P}^1$; (b) Affine charts of $\mathbb{R}\mathrm{P}^2$.

Example (m-dimensional projective space). To construct a chart of $\mathbb{R}\mathrm{P}^m$, fix a coordinate system $\{x_1, \dots, x_{m+1}\}$ of $\mathbb{R}^{m+1}$. Since every line through the origin in $\mathbb{R}^{m+1}$, not lying in the hyperplane $x_1 = 0$, intersects the hyperplane $x_1 = 1$, the remaining coordinates $\lambda_1 = (x_2, \dots, x_{m+1}) \in \mathbb{R}^m_1$ on this affine hyperplane ($x_1 = 1$) parametrise 'almost all' points of the projective space (Fig. 1.6). The missing lines of this chart, contained in the hyperplane $x_1 = 0$, correspond to "infinitely far points" of the affine hyperplane $x = 1$.

These missing lines are the points of a projective space of dimension $m - 1$ ("at infinity" for this chart). Therefore $\mathbb{R}\mathrm{P}^m = \mathbb{R}^m \sqcup \mathbb{R}\mathrm{P}^{m-1}$. For example, the projective plane $\mathbb{R}\mathrm{P}^2$ is the disjoint union of an affine plane and a projective line "at infinity". Conversely, the complement of any projective hyperplane $\mathbb{R}\mathrm{P}^{m-1}$ of $\mathbb{R}\mathrm{P}^m$ is an affine chart $\mathbb{R}^m$ of $\mathbb{R}\mathrm{P}^m$.

We construct other charts by fixing any other hyperplane $x_k = 1$. It parametrise again 'almost all' points of the projective space:

$$\lambda_k = (x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_{m+1}) \in \mathbb{R}^m_k \ .$$

Taking k from 1 to $m + 1$, we obtain $m + 1$ charts covering the whole projective space:

$$\mathbb{R}\mathrm{P}^m = \mathbb{R}^m_1 \cup \mathbb{R}^m_2 \cup \cdots \cup \mathbb{R}^m_{m+1}$$

(here the unions are not disjoint). This shows that $\mathbb{R}\mathrm{P}^m$ is covered by $m+1$ such charts. In fact the following holds:

Theorem. *The projective space* $\mathbb{R}\mathrm{P}^m$ *is a connected closed smooth manifold.*

EXERCISE. An embedding of $\mathbb{R}\mathrm{P}^2$ into $\mathbb{R}^4$ is provided on page 24. We left to the reader to construct the analogous embeddings $\mathbb{R}\mathrm{P}^m \to \mathbb{R}^N$.

Homogeneous Coordinates on $\mathbb{R}\mathrm{P}^m$. One associates the *homogeneous coordinates* $[x_0 : \cdots : x_m]$ to the element of $\mathbb{R}\mathrm{P}^m$ which represents the line of $\mathbb{R}^{m+1}$ joining the origin to the point $(x_0, \dots, x_m)$ different from the origin. Thus, the homogeneous coordinates $[\lambda x_0 : \cdots : \lambda x_m]$ represent the same point of the projective space if $\lambda \neq 0$.

A polynomial P_k of $n + 1$ variables is *homogeneous of degree* k if all its monomials have total degree k. Observe that if a point $(x_0, \dots, x_n) \neq O$ of $\mathbb{R}^{n+1}$ satisfies a homogeneous equation $P_k = 0$, then all points of the line through the origin $\{(\lambda x_0, \dots, \lambda x_n) : \lambda \in \mathbb{R}\}$ also satisfy that equation.

For example, if a point $(x, y, z) = (a, b, c)$ belongs to the surface given by the homogeneous equation $x^4 + xyz^2 + y^3 z = 0$, the line $\{(\lambda a, \lambda b, \lambda c) : \lambda \in \mathbb{R}\}$ is contained in the surface because its points satisfy the equation:

$$(\lambda a)^4 + (\lambda a)(\lambda b)(\lambda c)^2 + (\lambda b)^3(\lambda c) = \lambda^4(a^4 + abc^2 + b^3 c) = 0.$$

This line is the point of $\mathbb{R}\mathrm{P}^2$ of homogeneous coordinates $[a : b : c]$. All such lines of our surface in $\mathbb{R}^3$ are points of $\mathbb{R}\mathrm{P}^2$ forming a curve whose equation (in homogeneous coordinates $[x : y : z]$) is also $x^4 + xyz^2 + y^3 z = 0$.

EXERCISE (Conics). Let $\mathcal{C}$ be the curve in $\mathbb{R}\mathrm{P}^2$ determined by the cone in $\mathbb{R}^3$ of homogeneous equation $z^2 = 2x^2 + y^2$. Show three affine charts of $\mathbb{R}\mathrm{P}^2$ in which the curve $\mathcal{C}$ appears respectively as ellipse, hyperbola and parabola.

SOLUTION. e) The intersection of the cone with the plane $z = 1$ (our first affine chart) is the ellipse of equation $2x^2 + y^2 = 1$.

h) In the chart given by the plane $x = 1$ we get the hyperbola $z^2 - y^2 = 2$.

p) The cone intersects the plane $z = y + 1$ (our third affine chart) along a parabola whose projection to the xy-plane has equation $y = x^2 - \frac{1}{2}$. Verify it !

Therefore these ellipse, hyperbola and parabola are just three different representatives (in our three respective affine charts) of the curve $\mathcal{C}$ in $\mathbb{R}\mathrm{P}^2$.

Thus the ellipse, hyperbola and parabola, which were topologically different in Euclidean plane, become indistinguishable in the projective plane.

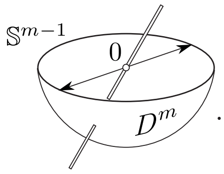
Other constructions We show now a construction of the projective space $\mathbb{R}\mathrm{P}^m$ that generalise the angle coordinate φ described above for $\mathbb{R}\mathrm{P}^1$.

Every line passing through the origin of $\mathbb{R}^{m+1}$ intersects the sphere $\mathbb{S}^m \subset \mathbb{R}^{m+1}$ at two antipodal points. Hence, the projective space $\mathbb{R}\mathrm{P}^m$ can be obtained from the sphere $\mathbb{S}^m$ by identifying its antipodal points:

$$\mathbb{R}\mathrm{P}^m = \frac{\mathbb{S}^m}{\mathbb{S}^0}\,, \qquad \text{where } \ \mathbb{S}^0 = \{\pm 1\}.$$

One can also consider only one hemisphere of the sphere (which is topologically a disc D^m). Next one has to identify the opposite points on its boundary $\partial D^m = \mathbb{S}^{m-1}$. Thus,

$$\mathbb{R}\mathrm{P}^m = \frac{D^m}{\pm \text{ on } (\partial D^m = \mathbb{S}^{m-1})} \quad :$$



.

Exercise. Show that $\mathbb{R}\mathrm{P}^2$ is not orientable. Hint: Show that the complement of a disc in $\mathbb{R}\mathrm{P}^2$ is a Möbius band.

A manifold is *simply connected* if every closed path in it can be continuously contracted to a point.

Exercise. Prove that the real projective plane $\mathbb{R}\mathrm{P}^2$ is not simply connected.

For the classification of the non-orientable surfaces, the projective plane plays a role similar to that of the sphere for orientable surfaces. The following theorem provides the complete classification of the closed smooth surfaces.

Theorem. *Every connected closed smooth surface is diffeomorphic either to a sphere with g handles or to a real projective plane with g handles.*

Example. The real projective plane with one handle is called *Klein Bottle*. The Klein Bottle can be also constructed from a cylinder by gluing the opposite boundaries as shown in Fig. 1.7.

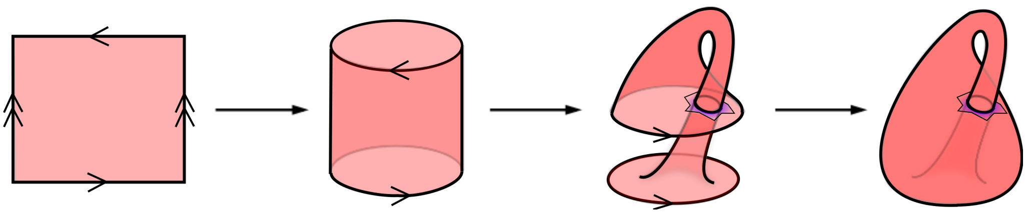


Figure 1.7: Klein Bottle construction.

Remark. In contrast to the complete classification of the real smooth closed surfaces, the classification of 3-dimensional manifolds is not known!

Working at his Urbino Castle office, Uccello made a lot of architectural projects by using his studies on projective geometry. Vasari mentioned that Uccello fanatically worked into the night and did not come to the bed when his wife called him, usually replying, as if speaking about his mistress: "*Oh, what sweet thing is this perspective!*".

Exercises. **1**. Is $\mathbb{R}\mathrm{P}^n$ orientable? [*Hint*: the answer depends on the dimension n.]

2. Compute the dimension of the *Grassmannian* manifold $G_k(\mathbb{R}^n)$, formed by all the k-dimensional vector subspaces of $\mathbb{R}^n$.

3. Compute the dimension of the manifold whose points are all the affine lines of $\mathbb{R}^n$ and of the manifold whose points are all the tangent lines of the sphere $\mathbb{S}^n$. Are these two manifolds diffeomorphic?

Complex projective space. The *complex projective space* $\mathbb{C}\mathrm{P}^n$ is the manifold whose points are the complex lines of $\mathbb{C}^{n+1}$ passing through the origin:

$$\mathbb{C}\mathrm{P}^n = \frac{\mathbb{C}^{n+1} \smallsetminus \{0\}}{\mathbb{C} \smallsetminus \{0\}} \ .$$

4. Prove that the complex projective line $\mathbb{C}\mathrm{P}^1$ is a connected closed smooth 2-manifold. Is it an orientable surface ? Which surface is it ?

5. Is $\mathbb{C}\mathrm{P}^n$ smooth? Orientable? Connected?

6. The elliptic curve. Consider d generic complex projective lines in $\mathbb{C}\mathrm{P}^2$. The affine equation of the union of these lines is $f = 0$, where f is a polynomial of degree d. For a generic small enough ε, the equation $f = \varepsilon$ defines a submanifold $V^1 \subset \mathbb{C}\mathrm{P}^2$ of real dimension 2. Show that V^1 is a sphere with g handles and calculate g.
(Hint. The solution is related to $\int dx/\sqrt{p_d(x)}$, where p_d is a polynomial of degree d.) The solution of this exercise is discussed below (pages 147-149).

1.4 The Tangent Space

Let M^k and N^ℓ be two smooth manifolds, of dimension k and ℓ respectively. A map $f : M^k \to N^\ell$ is said to be *smooth* if for every point $x \in M^k$ there exist two smooth coordinate systems $\{x_1, \ldots, k_k\}$ and $\{y_1, \ldots, y_\ell\}$ at x and at $f(x)$ for which f is smooth (i.e., all the components f_j are smooth). This definition is independent of the choice of the smooth coordinate systems.

Consider a smooth k-dimensional manifold M^k embedded in Euclidean space $\mathbb{R}^N$. The velocity vector of a smooth curve $\varphi : I \to M$, at the point $x \in M$, is defined as

$$\lim_{t \to 0} \frac{\varphi(t) - \varphi(0)}{t} \ , \qquad \text{where} \ \ \varphi(0) = x \ .$$

(Without the embedding of M in a vector space, the difference $\varphi(x) - \varphi(0)$ has no sense.)

Definition. The velocity vector of a curve in M emanating from x is called a *tangent vector* to M at x. The set $T_x M$ of all tangent vectors to M at x is called the *tangent space* to M at x – Fig.1.8.

The definition of tangent vectors can also be given in intrinsic way, independent of the embedding of M into $\mathbb{R}^N$.

Two curves $\varphi : I \to M$, $\psi : I \to M$ are *equivalent* (at x) if

$$\varphi(0) = \psi(0) = x \quad \text{and} \quad \lim_{t \to 0} \frac{\varphi(t) - \psi(t)}{t} = 0$$

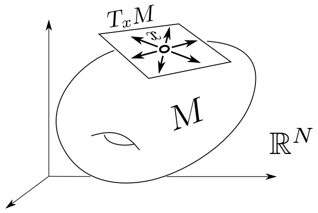


Figure 1.8: The tangent space.

in some chart.

Notice that equivalence of curves in a chart f (of the atlas) implies equivalence on any other chart g (since the transition map $g^{-1} \circ f$, from one chart to the other, is a diffeomorphism).

Definition. A *tangent vector* of a manifold M at a point x is an equivalence class of curves emanating from x $(\varphi : I \to M,\ \varphi(0) = x)$.

For embedded manifolds this definition agrees with the previous definition. Its advantage lies in the fact that it also holds for abstract manifolds, not embedded anywhere.

Theorem. *The set T_xM of tangent vectors of a k-dimensional manifold M at a given point x is a vector space of dimension k.*

Proof. The equivalence class of a curve φ is defined (for a fixed coordinate system) by the components of the velocity vector of $\varphi(t)$ at $\varphi(0)$.

Hence, our abstract vector defined without use of coordinates becomes an 'ordinary arrow' as soon as a coordinate system is fixed. We need only to prove the independence of the vector operations (addition and multiplication by scalars) from the coordinate system occurring in their definition. This independence follows from the linearity of the operation of taking the derivative of any "change of variables":

Fix a smooth coordinate system $\{x_1, \ldots, x_k\}$ at x. A curve φ is then given by k smooth functions $x_1 = \varphi_1(t), \ldots, x_k = \varphi_k(t)$. The components of the velocity vector v of φ at x are then defined by

$$v_1 = \frac{d\varphi_1}{dt}\,|_{t=0}\ , \ldots,\ v_k = \frac{d\varphi_k}{dt}\,|_{t=0}\ .$$

Now, by the definition of velocity, every curve ψ equivalent to φ at x satisfies:

$$v_1 = \frac{d\psi_1}{dt}\,|_{t=0}\ , \ldots,\ v_k = \frac{d\psi_k}{dt}\,|_{t=0}\ .$$

Therefore, $T_xM \simeq \mathbb{R}^k$. We shall prove that for any other smooth coordinate system $\{z_1, \ldots, z_k\}$, we get the same vector space. Let

$$z_1 = F_1(x_1, \ldots, x_k),\ \ldots,\ z_k = F_k(x_1, \ldots, x_k)$$

be a change of coordinates. In the new coordinate system, the components of the velocity vector v are given by

$$w_j = \frac{dz_j(\varphi(t))}{dt}\ |_{t=0} = \sum_{m=1}^{k} \frac{\partial F_j}{\partial x_m}\ v_m\ .^*$$

These linear equations state a linear relation between the velocities v and w, $w = Av$, where A is the *Jacobian matrix* of the coordinate change F:

$$(A)_{j,m} = \frac{\partial F_j}{\partial x_m}\ .$$

Consequently, the new coordinate system defines the same vector space structure on T_xM as the initial one. □

Example. Let M^n be a submanifold of an affine space $\mathbb{R}^N$. Then T_xM^n can be thought as an n-dimensional plane in $\mathbb{R}^N$ passing through x. However, in doing this one must keep in mind that *the tangent spaces to M at different points x and y are disjoint*: $T_xM^n \cap T_yM^n = \varnothing$ (one consists of vectors attached at the point x and the other of vectors attached at the point y).

1.5 Derivative of a Map

Let $f : M^k \to N^\ell$ be a smooth map between two smooth manifolds M^k and N^ℓ. A curve φ in M through x, with velocity $v \in T_xM$, induces, via f, a curve $\psi := f \circ \varphi$ in N (see Fig. 1.9-a) whose velocity vector at $f(x)$ is

$$w = \frac{d}{dt}\ |_{t=0}\ f(\varphi(t))\ .$$

---

*Physicists call *Einstein convention* the elimination of the sum symbol, writing usually

$$w_j = \frac{\partial F_j}{\partial x_m}\ v_m\,, \quad \text{instead of} \quad w_j = \sum_{m=1}^{k} \frac{\partial F_j}{\partial x_m}\ v_m\,.$$

Notice that the tangent vector $w \in T_{f(x)}N$ does not depend on the curve φ, it depends only on the vector v (and of course, on the point x and the map f). Indeed, if $\widetilde{\varphi}$ is another curve through x having the same velocity v as φ, then both induced curves ψ and $\widetilde{\psi} := f \circ \widetilde{\varphi}$ have the same velocity vector w at $f(x)$. Hence, we have constructed a map between the tangent spaces (see Fig. 1.9-b)

$$f_{*x} : T_xM \longrightarrow T_{f(x)}N \ , \quad f_{*x}v = w \ .$$

Definition. The map f_{*x} is the *derivative* of f at x. The vector $f_{*x}v$ is the *derivative of f at x in the direction of v.*

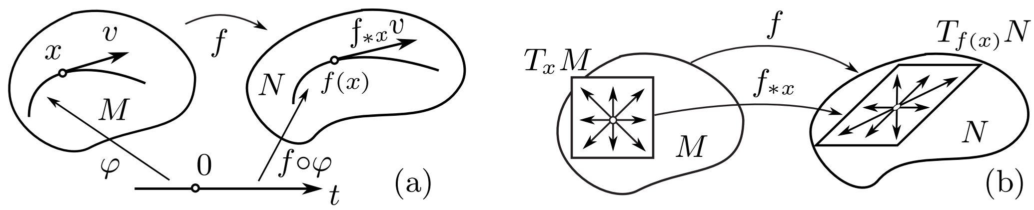


Figure 1.9: (a) The action of a smooth map on a vector; (b) The derivative of a map at a point.

If f is a function (that is, $N = \mathbb{R}$), then f_{*x} is also called the *differential* of f at x, and is also denoted df_x.

Problem. Prove that the map $f_{*x} : T_xM \to T_{f(x)}N$ is linear.

Solution. In local coordinates, f is given by ℓ smooth functions $f_1, \ldots, f_\ell$ of k variables, and the vectors $v = (v_1, \ldots, v_k)$ and $f_{*x}v = w = (w_1, \ldots, w_\ell)$ are related by the equation

$$w_j = \sum_{m=1}^{k} \frac{\partial f_j}{\partial x_m} \, v_m \ .$$

Once the local coordinate systems near x and $f(x)$ are fixed, the tangent spaces T_xM and $T_{f(x)}N$ can be identified respectively to $\mathbb{R}^k$ and $\mathbb{R}^\ell$, as explained above. Hence the map

$$f_{*x} : T_xM \longrightarrow T_{f(x)}N$$

is now identified to the linear operator $A : \mathbb{R}^k \longrightarrow \mathbb{R}^\ell$, whose matrix

$$(A)_{j,m} = \left(\frac{\partial f_j}{\partial x_m}\right)$$

is the Jacobian matrix of the map f in the considered local coordinates.

Problem. Consider the map of the line into the plane $f(x) = (\cos x, 2\sin x)$. Find the value of its derivative on the vector v of length 7 attached at the point $x_0 \in \mathbb{R}$ and orienting positively the x-axis.

Answer. $f_{*x_0} v = (-7\sin x_0, 14\cos x_0)$.

Besides the map $f : M \to N$, consider a manifold K, and a smooth map $g : N \to K$. The composition of f and g provides a map $h : M \to K$, denoted by $g \circ f$, as shown by the following (commutative) diagram:

$$\begin{tikzcd} M \arrow[r, "f"] \arrow[dr, "h=g\circ f"'] & N \arrow[d, "g"] \\ & K \end{tikzcd}$$

For each $x \in M$ a similar diagram of linear maps is induced for the derivatives:

$$\begin{tikzcd} T_xM \arrow[r, "f_{*x}"] \arrow[dr, "h_{*x}"'] & T_{f(x)}N \arrow[d, "g_{*f(x)}"] \\ & T_{h(x)}K \end{tikzcd}$$

Proposition (Chain rule)**.** *The derivative of the composition of two maps is the composition of their derivatives:*

$$(g \circ f)_{*x} = g_{*f(x)} \circ f_{*x} .$$

1.6 Critical Points of Maps

Recall that for a function $f : \mathbb{R}^n \to \mathbb{R}$, a point $x \in \mathbb{R}^n$ is called *critical* if the differential of f at x vanishes (i.e. if it is the zero map: $df(v) = 0$ for any $v \in T_x\mathbb{R}^n$). This notion can be extended to maps between manifolds.

Critical and Regular Points. Let $f : M \to N$ be a smooth map. A point $x \in M$ is a *critical point* of f if the derivative map of f at x is not surjective:

$$f_{*x}(T_x M) \neq T_{f(x)} N \ .$$

A point $y \in N$ is a *critical value* of f if its preimage $f^{-1}(y)$ contains a critical point. A point $x \in M$ (value $y \in N$) is said to be *regular* if it is not critical.

Problem. Find the set of all critical points of the map $f : (x, y) \mapsto (X, Y)$ of the plane to the plane $f(x, y) = (x^3 + xy, y)$ and find the image of this set under the map f. (This map is called *Whitney map* or *Whitney cusp*.)

Solution. Since f is a map between two manifolds of the same dimension ($M = \mathbb{R}^2$, $N = \mathbb{R}^2$), we have to look for the points $p = (x, y)$ at which the linear operator f_{*p} is not a linear isomorphism. Since the matrix of the derivative map is

$$(f_{*p}) = \begin{pmatrix} 3x^2 + y & x \\ 0 & 1 \end{pmatrix},$$

the derivative is degenerate on the parabola $y = -3x^2$. Its image under f is the semicubic parabola $(X/2)^2 + (Y/3)^3 = 0$ (verify it and draw the pictures!). This curve has a singular point (a cusp) at the origin.

Problem. Find the kernel of the derivative f_{*p} of the Whitney map at its critical points.

Solution. The derivative f_{*p} is a linear isomorphism at all points apart from the critical ones (which form the parabola $3x^2 + y = 0$). At those points the rank of the Jacobian matrix equals 1, and hence the kernel of the derivative has dimension 1.

Therefore there is a field of lines on the parabola of critical points, the kernel field of the derivative. At each point this kernel is parallel to the x-axis, since it is generated by the vector $v \in T_p\mathbb{R}^2$, $v = (1, 0)$. – Fig. 1.10.

To visualise better the Whitney map, we realise it as the projection of a smooth surface S in the 3-dimensional space to the horizontal plane.

For take the map $i : (x, y) \mapsto (X, Y, Z)$ of the plane into the 3-space, $i(x, y) = (\, x^3 + xy,\, y,\, x\,)$. The reader can verify that for any $p \in \mathbb{R}^2$ the Jacobian matrix (i_{*p}) has rank 2. Thus the map i embeds $\mathbb{R}^2$ into $\mathbb{R}^3$ as the smooth surface $S = i(\mathbb{R}^2)$. In particular, for any $p \in \mathbb{R}^2$ the derivative map i_{*p} sends the tangent plane $T_p\mathbb{R}^2$ isomorphically onto the tangent plane to

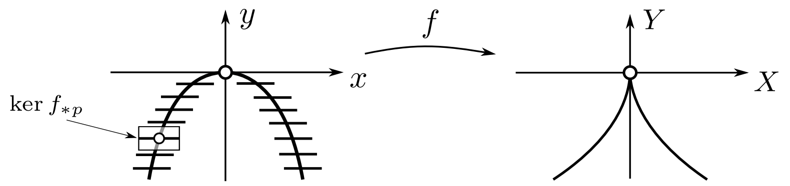


Figure 1.10: Critical points and critical values of the Whitney map.

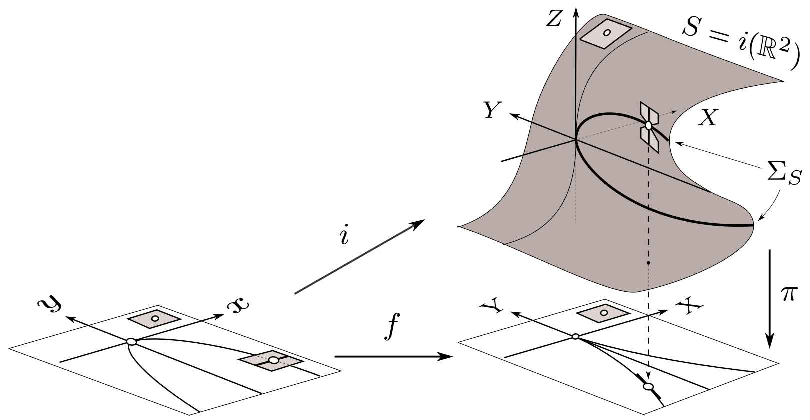


Figure 1.11: The Whitney map as the projection of a smooth surface.

S at $i(p)$ – Fig. 1.11. Now, project S to the horizontal plane by the vertical projection $\pi : (X, Y, Z) \mapsto (X, Y)$. Clearly $f = \pi \circ i$.

The critical points of the projection are those points of our surface S where the tangent plane is vertical. These points form a smooth curve Σ_S on S (it is the image under i of the parabola $y = -3x^2$, calculated above).

Remark. From the solution of the problem it is clear that the kernel of the derivative is tangent to the parabola of critical points only at the origin. H. Whitney proved that this type of "behaviour" is typical for smooth maps of the plane into the plane. For example, any smooth map $\widehat{f}$ near f has a similar "behaviour" near the origin: Its critical points form a smooth curve, and the kernel of the derivative is transverse to that curve at all these points excepted one (near the origin) at which the kernel is tangent. The image of this point under $\widehat{f}$ is a cusp of the curve of critical values.

Exercise. Consider the map of the plane of the complex variable $z = x + iy$ to the plane of the complex variable $w = X + iY$, given by the formula $w = z^2 + 2\overline{z}$, as a smooth map from the two-dimensional real plane to the two-dimensional real plane:

$$X = x^2 - y^2 + 2x, \qquad Y = 2xy - 2y.$$

A very instructive exercise is to find the set of all critical points of this map, the kernel of the derivative on the curve of critical points and the set of critical values, and draw the corresponding pictures.
Hint: The calculations are easier in the complex variables (for instance, the derivative of our map takes the value $dw(\xi) = 2z\xi + 2\overline{\xi}$ on the vector ξ).

1.7 Implicit Function Theorem

We start with an example. Consider a function $f : I \subset \mathbb{R} \to \mathbb{R}$, where I is a closed bounded interval such that $f(I) \subset I$. Suppose there exists $\theta \in \mathbb{R}$ such that $| f'(x) | < \theta < 1$. Then one can show that the graph of f, $y = f(x)$, has exactly one point of intersection with the diagonal $y = x$.

In order to generalise this fact, we remark that the mean value theorem implies that $| f(x) - f(y) | \leqslant \theta | x - y |$ for every $x, y \in I$.

Contraction. A map $f : \mathbb{R}^n \to \mathbb{R}^n$ is called a *contraction* if there exists $0 < \theta < 1$ such that

$$|| f(x) - f(y) || \leqslant \theta || x - y || \qquad \forall x, y \in \mathbb{R}^n.$$

Theorem. *Every contraction map on $\mathbb{R}^n$ has exactly one fixed point.*[*]

Exercise. Prove the above theorem. Hint: Given a contraction map f and a point x, consider the sequence defined by $x_1 = x$ and $x_{n+1} = f(x_n)$. Then show that: a) the sequence is convergent; b) its limit is a fixed point of f; and c) there is no other fixed point of f.

Exercise. Is the map $f(x) = \sqrt{x^2 + 1}$ a contraction?

We suppose now that this construction depends on a parameter: namely, we consider the intersections of the graph $y = f(x)$ with a family of lines parallel to the diagonal $y = x$. Since the graph $y = f(x)$ has exactly one intersection with the diagonal, the graph has also exactly one point of intersection with every line parallel to $y = x$, provided that this line is sufficiently close to the diagonal.

We can reduce this situation by a coordinate change to the case of the intersections of the graph $y = f(x)$ of a function with the horizontal lines

[*] The definition of contraction extends to any metric space, just replacing the Euclidean distance $||b - a||$ by the metric $\varrho(a, b)$. This theorem holds on any complete metric space.

$y =$ const. In this case, if f' never vanishes, then for every y the horizontal line $y =$ const intersects the graph $y = f(x)$ exactly at one point whose x-coordinate is a smooth function of y: $x = g(y)$.

We get the inverse function theorem in the unidimensional case:

Proposition. *Let $f : I \to \mathbb{R}$ be a smooth function defined on a compact interval. Suppose that $f'(x) \neq 0$ for every $x \in I$. Then, there exists a smooth function $x = g(y)$ such that $g \circ f(x) = x$.*

Exercise. Show that the derivative of this function is $g'(y)|_{y=f(x)} = 1/f'(x)$.

The above ideas can be generalised to maps:

Implicit Function Theorem ($\mathbb{R}^2 \to \mathbb{R}$)**.** *Let $(x_0, y_0) \in \mathbb{R}^2$ be a zero of a smooth map $F : \mathbb{R}^2 \to \mathbb{R} : F(x_0, y_0) = 0$. If $\frac{\partial F}{\partial x}(x_0, y_0) \neq 0$, then the set $\{(x, y) \in \mathbb{R}^2 : F(x, y) = 0\}$ can be expressed, near (x_0, y_0), as the graph of a function $x = g(y)$, that is, $F(y, g(y)) = 0$.*

Indeed, using the condition $\frac{\partial F}{\partial x}(x_0, y_0) \neq 0$ together with the above reasoning one gets that for any y sufficiently close to y_0 the horizontal line $y =$ const intersects the set defined by the equation $F(x, y) = 0$ exactly at one point whose x-coordinate is a smooth function of y: $x = g(y)$ – Fig. 1.12.

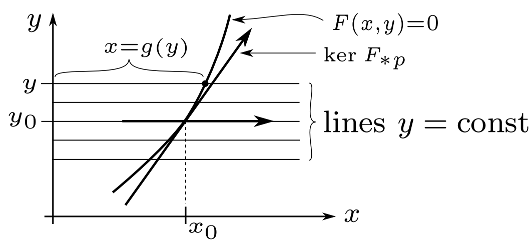


Figure 1.12: Implicit function theorem.

For a general map $F : \mathbb{R}^{k+\ell} \to \mathbb{R}^k$ the crucial condition on the derivative is replaced by a suitable condition on the Jacobian matrix of that map.

Equip the space $\mathbb{R}^{k+\ell}$ with coordinates $\{x, y\}$, where $x = (x_1, \ldots, x_k)$ and $y = (y_1, \ldots, y_\ell)$. We need to consider the matrix

$$\frac{\partial F}{\partial x} := \left(\frac{\partial F_i}{\partial x_j}\right)_{i,j=1,\ldots,k}.$$

Theorem (Implicit Function Theorem)**.** *Suppose that $F(x_0, y_0) = 0$ and*

$$\det \frac{\partial F}{\partial x}(x_0, y_0) \neq 0 \ .$$

Then, near the point (x_0, y_0), the set $\{(x, y) : F(x, y) = 0\}$ is the graph $\{x = g(y)\}$ of a smooth map g.

EXERCISE. Prove that in the case of an affine map $F : \mathbb{R}^{k+\ell} \to \mathbb{R}^k$ the theorem holds globally. Indeed, the projection along the k-planes $y =$ const of the ℓ-space $\mathbb{R}^\ell_y$ to the ℓ-dimensional affine subspace $F = 0$ is an isomorphism of affine spaces: The k-planes $y =$ const intersect each of these two spaces at one point – the projection sends one point to the other – Fig 1.13.

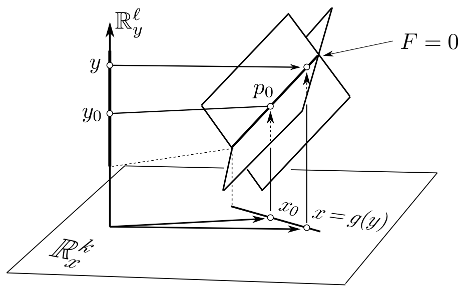


Figure 1.13: Implicit function theorem (affine case).

1.8 Critical Values of Smooth Maps

The set of critical points of a map, called *critical set*, can be very complicated. For example, *every closed subset of a manifold can be viewed as the critical set of a suitable map.* However, there is an important result (proved in Section 1.9) for the set of critical values of a map, known as *Sard Lemma*:

Theorem (Bertini-Sard-Morse)**.** *The set of critical values of a smooth map $f : M \to N$ has Lebesgue measure zero.*

Measure Zero. A subset of $\mathbb{R}^k$ has *measure zero* if for any $\varepsilon > 0$ it is possible to cover it by a sequence of cubes of $\mathbb{R}^k$ having total k-dimensional volume less than ε. A subset C of a k-dimensional manifold N has measure zero if for any coordinate system $\psi : W \subset N \to \mathbb{R}^k$ the set $\psi(W \cap C)$ has measure zero in $\mathbb{R}^k$.

An important consequence of this Theorem is the following result.

Corollary (Gibbs Principle). *Every smooth n-dimensional manifold can be embedded (resp. immersed) into $\mathbb{R}^K$ for any $K \geqslant 2n+1$ (resp. $K \geqslant 2n$).*

Proof. Let M^n be a manifold embedded into some $\mathbb{R}^N$, with $N > 2n+1$. Consider a projection $\mathbb{R}^N \to \mathbb{R}^{N-1}$.

The restriction of this projection to M^n is not an embedding if: (1) the direction of the projection is tangent to M^n at some point; or (2) the direction of the projection is parallel to a chord of the manifold M^n (see Fig. 1.14).

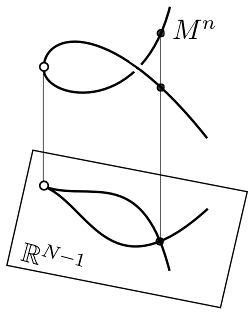


Figure 1.14: Singularities of projections of a manifold.

Now, on the one hand, we have that the set $\mathcal{L}$ whose elements are all the lines tangent to M^n is a manifold of dimension $2n-1$, and the set $\mathcal{C}$ formed by the chords of M^n is a manifold of dimension $2n$.

On the other hand, the space of all parallel projections $\mathbb{R}^N \to \mathbb{R}^{N-1}$ is of dimension $N-1$ (each direction of projection is a point of $\mathbb{R}\mathrm{P}^{N-1}$).

If $N-1 > 2n$ we can find a projection $\mathbb{R}^N \to \mathbb{R}^{N-1}$ such that the above two cases never produce (by the Sard Lemma, there exist a point of $\mathbb{R}\mathrm{P}^{N-1}$ which does not belong to the image of the natural map $\mathcal{L} \cup \mathcal{C} \to \mathbb{R}\mathrm{P}^{N-1}$).

Consequently, the restriction to M^n of that projection $\mathbb{R}^N \to \mathbb{R}^{N-1}$ is an embedding of M^n into $\mathbb{R}^K = \mathbb{R}^{N-1}$. Thus, M^n can be embedded into $\mathbb{R}^K$, if $K > 2n$, that is, if $K \geqslant 2n+1$. $\square$

Example. There exists an embedding of $\mathbb{R}\mathrm{P}^2$ into $\mathbb{R}^5$. Moreover, there exists an embedding of $\mathbb{R}\mathrm{P}^2$ into $\mathbb{R}^4$. Hint: One can embed $\mathbb{R}\mathrm{P}^2$ into $\mathbb{R}^5$ in such a way that the image lies on the sphere $\mathbb{S}^4$.

1.9 Proof of Sard Lemma

Although we have stated Sard Lemma for two arbitrary smooth manifolds M^n and N^k, the Remark of page 20 permit us to reduce the theorem to the case in which N is a Euclidean space $\mathbb{R}^k$ and f is a map of the n-cube $I^n = [0,1] \times \cdots \times [0,1]$ to $\mathbb{R}^k$, $f : I^n \to \mathbb{R}^k$.

1st Case. Consider a real function $f : I \to \mathbb{R}$. Assuming f'' continuous on I, there exists $C > 0$, such that $|f''(x)| < C \quad \forall x \in I$. Then the mean value theorem (for f') implies $|f'(x) - f'(x_0)| < C|x - x_0|$, $\forall x, x_0 \in I$.

Subdivide the interval I into n segments of equal length $1/n$. The above inequality implies that for a segment containing a critical point x_0, the absolute value of the derivative is bounded above by C/n: $\quad |f'(x)| \leqslant C/n$. Therefore, the length of the image under f of such a segment is bounded above by C/n^2. Hence the sum of the lengths of the images of those segments does not exceed $n \cdot (C/n^2) = C/n$. As $n \to \infty$ this sum of lengths tends to zero. Consequently, the measure of the set of critical values is zero.

2th Case. We consider now a function of two variables x and y, defined on the square $I^2 = [0,1] \times [0,1]$. At the critical points, the two first order partial derivatives $\partial_x f$ and $\partial_y f$ are zero. Consider the curve of equation $\partial_x f(x,y) = 0$. At its singular points, the second order partial derivatives

$$\frac{\partial^2 f}{\partial x^2}(x,y) \quad \text{and} \quad \frac{\partial^2 f}{\partial x \partial y}(x,y)$$

both vanish. Therefore, the set formed by the critical values of f attained at the smooth part of the curve $\partial_x f(x,y) = 0$ has measure zero, by the 1st case. By the same arguments, the measure of the set formed by the critical values of f attained at the smooth part of the curve $\partial_y f(x,y) = 0$ is zero.

It remains to look at the critical points of f such that

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y^2} = 0\ .$$

As above, divide the square I^2 into n^2 sub-squares of equal area $1/n^2$.

Assuming f of class C^3, for any two points p, q of a sub-square containing such a critical point, we have $|f(p) - f(q)| \leqslant C/n^3$, where C is a constant independent of that sub-square and n (indeed, the quadratic part of f is zero at these critical points). Hence the length of the image of such a sub-square is bounded above by C/n^3. Since the total number of such sub-squares is at most n^2, the sum of the measures of their images is bounded above by $n^2 \cdot C/n^3 = C/n$. Thus the measure of the image of these critical points is zero.

We have proved that the set of critical values of f has measure zero, since it is the union of three subsets, each one of measure zero.

3th Case. Finally, consider a map of the square $I^2 = [0,1] \times [0,1]$ to the plane $\mathbb{R}^2$:

$$X = f_1(x,y) \ , \quad Y = f_2(x,y) \ .$$

Let (x_0, y_0) be a critical point of f. If $\frac{\partial f_1}{\partial x}(x_0,y_0) \neq 0$, then the curve $f_1(x,y) = f_1(x_0,y_0)$ is smooth (regular) near (x_0,y_0), due to the Implicit Function Theorem. Take $\widetilde{x} := f_1$ as new local coordinate replacing x. In this new coordinate system, f is expressed near the critical point (x_0,y_0) as

$$X = \widetilde{x} \ , \qquad Y = f_2(x(\widetilde{x},y),y) =: F(\widetilde{x},y) \ ,$$

where F is a smooth function. That is, f is expressed as a 1-parameter family of smooth functions of the variable y. On every fibre $X = \text{const}$, the set of critical values has measure zero (again by the 1st case). By the Fubini theorem, the measure of the total set of such critical values is also zero.

The corresponding result for the second component function f_2 is obtained in the same way.

Hence, it remains to consider the critical points at which the Jacobian matrix of f is the zero matrix. Divide I^2 into n sub-squares of edge $1/n$. The air of the image of each sub-square containing one of those critical points is bounded above by C/n^4, C being independent of the sub-square (this estimation follows from the condition $Df = 0$ at the critical point). Since the total number of such sub-squares is at most n^2, the measure of the set of critical values corresponding to these critical points can not exceed C/n^2, so it is zero (since this holds for arbitrary n).

Exercise. Prove the Sard Lemma for the general case of a map $f : \mathbb{R}^n \to \mathbb{R}^p$. Hint: Adapt the above arguments.

Remark. If $\dim M \leqslant \dim N$, then every point of M is critical, so one has to prove that $f(M)$ has measure zero.

Remark. If the condition on the required smoothness of the map is not fulfilled, the Theorem does not hold. However, such counter-examples are not easy to construct.

1.10 Embedding of $\mathbb{R}\mathrm{P}^2$ into $\mathbb{R}^4$

Immersion and Embedding. A map $f : M \to N$ is an *immersion* if the derivative f_{*x} at any point $x \in M$ is injective (defining therefore an isomorphism of the tangent spaces T_xM and $T_{f(x)}f(M)$) and f is an *embedding* if moreover it is a homeomorphism onto its image.

Theorem. *There exists an embedding of* $\mathbb{R}\mathrm{P}^2$ *into* $\mathbb{R}^4$.

Proof. Since $\mathbb{R}\mathrm{P}^2 = \mathbb{S}^2/\{\pm 1\}$, any function on $\mathbb{R}\mathrm{P}^2$ is defined by an "even function" on the sphere. So to embed $\mathbb{R}\mathrm{P}^2$ into some $\mathbb{R}^N$ we use "even functions" on the unit sphere $x^2+y^2+z^2=1$. The quadratic forms suffice.

Consider the spaces $\mathbb{R}^3$ and $\mathbb{R}^6$, equipped respectively with coordinates $\{x,y,z\}$ and $\{u,v,w,p,q,r\}$. Consider the map $f : \mathbb{S}^2 \to \mathbb{R}^6$ defined by

$$f(x,y,z) := (x^2, y^2, z^2, xy, yz, xz) \ .$$

Exercise. Show that f is an immersion of the unit sphere $\mathbb{S}^2$ into $\mathbb{R}^6$.

Since $f(x,y,z) = f(-x,-y,-z)$, the immersion f induces a map $\widetilde{f}$ on the projective plane:

$$\begin{tikzcd} \mathbb{S}^2 \arrow[r, "f"] \arrow[d] & \mathbb{R}^6 \\ \mathbb{R}\mathrm{P}^2 \arrow[ur, "\widetilde{f}"'] & \end{tikzcd}$$

Exercise. Show that $\widetilde{f}$ is an embedding of $\mathbb{R}\mathrm{P}^2$ into $\mathbb{R}^6$.

Problem. Prove that the image $\widetilde{f}(\mathbb{R}\mathrm{P}^2)$ is contained in a hyperplane of $\mathbb{R}^6$.

Solution. It is contained in the hyperplane $\{u+v+w=1\} \simeq \mathbb{R}^5 \subset \mathbb{R}^6$ because $x^2+y^2+z^2$ is equal to 1 on the unit sphere.

Similarly, the image $\widetilde{f}(\mathbb{R}\mathrm{P}^2)$ is contained into an ellipsoid of this hyperplane. Indeed, the equality $(x^2+y^2+z^2)^2 = 1$ on the unit sphere implies that the corresponding image coordinates verify

$$u^2+v^2+w^2+2p^2+2q^2+2r^2 = 1 \ ,$$

which is the equation of an ellipsoid, whose intersection with the hyperplane is a 4-dimensional ellipsoid (diffeomorphic to $\mathbb{S}^4$). Now, by the Sard Lemma, there exists a point in the ellipsoid which does not belong to $\widetilde{f}(\mathbb{R}\mathrm{P}^2)$.

Applying a stereographic projection of the ellipsoid onto $\mathbb{R}^4$, from that point, we get an embedding of $\mathbb{R}\mathrm{P}^2$ into $\mathbb{R}^4$, proving the theorem. $\square$

Exercise. Show that there is no embedding of $\mathbb{R}\mathrm{P}^2$ into $\mathbb{R}^3$.

Exercise. Repeat the same construction for $\mathbb{R}\mathrm{P}^n$.
Hint: Noting that the space of *quadratic forms* on $\mathbb{R}^3$ is isomorphic to $\mathbb{R}^6$,

$$S^2\mathbb{R}^{3*} := \{(u\,x^2 + v\,y^2 + w\,z^2 + p\,xy + q\,yz + r\,zx)\} \simeq \mathbb{R}^6 \ ,$$

you can imitate the above map $f : \mathbb{S}^2 \to \mathbb{R}^6$ with the natural map

$$\mathbb{S}^n \to S^2\mathbb{R}^{(n+1)*} \simeq \mathbb{R}^{(n+2)(n+1)/2} \ .$$

In particular, for $n = 3$ one obtains an embedding $\mathbb{R}\mathrm{P}^3 \to \mathbb{S}^8 \subset \mathbb{R}^9 \subset \mathbb{R}^{10}$.

On the abstract definition of manifold. Claiming to generalise the "naïve" submanifolds in Euclidean space, mathematicians introduced an abstract definition of manifold. A *differentiable manifold* M, is a set M together with a finite or countable collection of charts $f : U \subset \mathbb{R}^k \to M$ covering it and satisfying the following conditions.

It is assumed that if for two charts f_i, f_j the sets $f_i(U_i)$ and $f_j(U_j)$ intersect, then the sets $V_i = f_i^{-1}(f_i(U_i) \cap f_j(U_j))$ and $V_j = f_j^{-1}(f_i(U_i) \cap f_j(U_j))$ are open. (The transition map from one chart to the other $f_{ij} := f_j^{-1} \circ f_i : V_i \to V_j$, called *coordinate change*, being then a map of open subsets of vector spaces.)

Two charts f_i and f_j are called *compatible* either if both coordinate changes f_{ij} and f_{ji} (which are defined if $f_i(U_i) \cap f_j(U_j)$ is nonempty) are differentiable or if $f_i(U_i) \cap f_j(U_j) = \varnothing$.

A collection of charts $f_i : U_i \to M$ is an *atlas* on M if every point of M is represented in at least one chart and if any two charts are compatible. Two atlases on M are *equivalent* if their union is again an atlas (i.e. if any chart of the first atlas is compatible with any chart of the second).

A *differentiable manifold structure* on M is a class of equivalent atlases.

To avoid pathologies, it is frequently supposed the differentiable structure has at least one atlas with a finite or countable set of charts and it is assumed that every two different points have non-intersecting neighbourhoods.
A *neighbourhood* of a point p on a manifold M is the image under a chart $f : U \subset \mathbb{R}^k \to M$ of a neighbourhood of the representation of that point $f^{-1}(p) \in U$ in that chart.

In this abstract view of manifolds, the realisability of any "abstract smooth manifold" by a smooth submanifold in a Euclidean space is called *Whitney Embedding Theorem*. This theorem shows that the abstract definition of manifold encompasses a set of objects no larger than the n-dimensional submanifolds in N-dimensional Euclidean space. Thus the axioms of the abstract notion of manifold add nothing to the elementary submanifolds in Euclidean spaces! So after Whitney Theorem, no longer was there any confusion as to if abstract manifolds (defined via charts) were any more general than submanifolds of Euclidean space. However, an advantage of using also the abstract approach is that it directly encompasses the cases when no embedding in Euclidean space is given in advance, as for the projective space, the universal covering of a manifold (defined in Ch. 3) or many other manifolds. In such cases, introducing an embedding would only lead to unnecessary complications that are avoided thanks to the Whitney Embedding Theorem. For example, the charts given on page 8 form an atlas of the projective space $\mathbb{R}\mathrm{P}^m$.

1.11 The Tangent Bundle

With each smooth manifold M there is associated another manifold whose dimension is twice the dimension of M.

Theorem. *The union of the tangent spaces to a manifold M at all its points, $TM = \cup_{x \in M} T_x M$, has a natural structure of smooth manifold.*

Proof. Local coordinates on TM are constructed as follows. Consider a chart $f : U \to M$ on the manifold M. Its inverse defines local coordinates $x_1, \dots, x_n$ in $W = f(U)$. Every vector ξ tangent to M at a point $x \in W$ is determined by its set of components $\xi_1, \dots, \xi_n$ in the given coordinate system. Specifically, if $\varphi : I \to M$ is a curve (in the class defining ξ) emanating from x at the instant $t = t_0$, then $\xi_i = \dfrac{d}{dt}|_{t=t_0} x_i(\varphi(t))$.

Hence every vector ξ tangent to M at a point of the domain W is determined by a set of $2n$ numbers $(x_1, \dots, x_n)$, $(\xi_1, \dots, \xi_n)$, the n coordinates of the "point of attachment" x and the n "components" ξ_i. We have thus constructed a coordinate system on a part of the set TM:

$$\psi : TW \to \mathbb{R}^{2n}, \qquad \psi(\xi) = (x_1, \dots, x_n, \xi_1, \dots, \xi_n),$$

whose inverse is a chart of a part of TM.

These charts of TM, obtained from the different charts of the atlas of M, are compatible*. Indeed, if $(y_1, \dots, y_n)$ is another local coordinate system on M and the components of the vector ξ in this coordinate system are $(\eta_1, \dots, \eta_n)$, then

$$y_i = y_i(x_1, \dots, x_n), \qquad \eta_i = \sum_{j=1}^{n} \frac{\partial y_i}{\partial x_j} \xi_j \qquad (i = 1, \dots, n),$$

are smooth functions of x_i and ξ_j.

In consequence the set TM of the vectors tangent to M has been provided with a structure of smooth manifold of dimension $2n$. □

*These charts of TM are of class C^{r-1} if M is of class C^r.

The manifold TM is called the *tangent bundle* of the manifold M.

Standard Projection and Sections. There are two natural maps between a manifold M and its tangent bundle TM – Fig. 1.15 :

The *null section* $i : M \to TM$ sends the point x to the zero vector of T_xM;

The *standard projection* $\pi : TM \to M$ sends the vector ξ to the point x at which ξ is tangent to M.

Exercise. Prove that the null section i and the standard projection π are smooth maps, that i is a diffeomorphism of M onto $i(M)$, and that $\pi \circ i : M \to M$ is the identity map – Fig. 1.15. Hint: Local coordinates are useful.

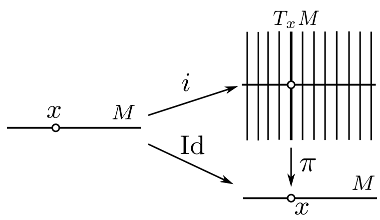


Figure 1.15: The tangent bundle.

The preimages of the points $x \in M$ under $\pi : TM \to M$ are called *fibres* of the bundle. Every fibre has the structure of a vector space (the fibre over the point x is the vector space T_xM). The manifold M is called the *base* of the bundle TM.

The tangent bundle is a special case of a vector bundle and of the more general concept of a fibre bundle. These concepts are fundamental in topology and analysis, and enable to carry over to manifolds the whole theory of ordinary differential equations.

Vector Fields. A *vector field* v (also called a *section of the tangent bundle*) on a smooth manifold M is a smooth map $v : M \to TM$ such that the composition $\pi \circ v : M \to M$ is the identity: The diagram

$$\begin{tikzcd} M \arrow[r, "v"] \arrow[dr, "\mathrm{Id}"'] & TM \arrow[d, "\pi"] \\ & M \end{tikzcd}$$

is commutative, that is, $\pi(v(x)) = x$.

Example. If M is a domain of the space $\mathbb{R}^n$, this definition coincides with the usual one: A *vector field* v is defined in M if to each point x there is assigned a vector $v(x) \in T_xM$ attached at that point and depending smoothly on x. Once a coordinate system is fixed on $\mathbb{R}^n$, the vector field v assigns to each point x an "arrow" tangent to M at x.

Parallelisable manifolds. The tangent bundle of a domain of the affine space $U \subset \mathbb{R}^n$ has a supplementary structure of direct product: $TU = U \times \mathbb{R}^n$. In fact, we can define a tangent vector to U by a pair (x, ξ), where $x \in U$ and ξ belongs to the vector space $\mathbb{R}^n$ for which an isomorphism with T_xU is defined. So, an equality is defined for tangent vectors at different points x and y of U and we say that the domain U (or the affine space) is *parallelised*.

The tangent bundle of a manifold is not necessarily a direct product and, in general, there is no reasonably natural way of defining the equality of vectors applied at different points of a manifold.

Definition. A manifold M is called *parallelised* if its tangent bundle is provided with a structure of direct product, that is, if there is given a diffeomorphism $TM^n = M \times \mathbb{R}^n$ transforming linearly T_xM in $x \times \mathbb{R}^n$. A manifold is said to be *parallelisable* if it is susceptible to be parallelised.

Exercise. Show that the cylinder is parallelisable, but the Möbius band is not (see Fig. 1.3). Show that the torus $\mathbb{T}^2$ is parallelisable.

We state the following theorem without proof:

Theorem. *Among the spheres $\mathbb{S}^n$, only $\mathbb{S}^1$, $\mathbb{S}^3$ and $\mathbb{S}^7$ are parallelisable.*

In particular, $\mathbb{S}^2$ is not parallelisable: $T\mathbb{S}^2 \neq \mathbb{S}^2 \times \mathbb{R}^2$. It follows, for example, that it is not possible to brush the 2-sphere (see Fig. 3, in Preface).

Exercise. Show that $\mathbb{S}^3$ is parallelisable. *Hint:* $\mathbb{S}^3$ is the group formed by the quaternions of length 1 (see Section 3.11).

The Tangent Map. Given a smooth map $f : M \to N$ from the manifold M to the manifold N, it induces for each $x \in M$ the linear map (the derivative of f at x):

$$f_{*x} : T_xM \to T_{f(x)}N.$$

Letting x range over M, the linear operator f_{*x} defines a map

$$f_* : TM \to TN, \qquad f_*|_{T_xM} = f_{*x},$$

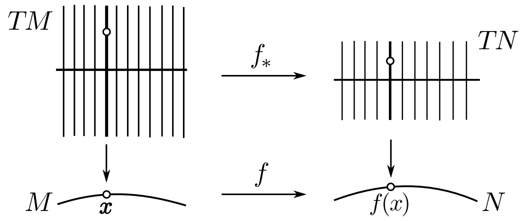


Figure 1.16: The tangent map.

of the tangent bundle of M to the tangent bundle of N. This map is smooth (prove it) and maps the fibres of TM linearly to the fibres of TN – Fig. 1.16.

The map f_* (noted also by $Tf : TM \to TN$) is called the *tangent map* to f.

Chapter 2

Group invariance and geometry

God always geometrises! Plato

The abstract notions of modern mathematics are difficult to follow and to understand if, as it is usual in modern universities, one starts from their axiomatic definitions rather than from the naive examples that preceded those definitions. Those examples are hidden from the students with the goal to enhance the authority of the professors. Very often the examples contain all the axiomatically defined abstract objects of the theory.

Elementary submanifolds in Euclidean spaces, like ellipses or hyperboloids, are easily understandable. The "abstract manifolds" are just the same elementary things considered, however, up to their diffeomorphisms in the smooth topology or up to their homeomorphisms in ordinary topology. In the smooth case, one makes from all ellipses the same "abstract manifold". One includes all simple closed polygons to the same "circular" topological manifold $\mathbb{S}^1$.

The proof of the realisability of any "abstract smooth manifold" by a smooth submanifold in a Euclidean space (*Whitney embedding theorem*) is a not too difficult exercise in the manipulation of the axioms listed in Remark of p. 25. The main consequence of this Whitney Theorem is that those axioms (of the abstract notion of manifold) add nothing to the elementary objects whose study lead to the axioms invention. It proves the axioms completeness rather than any property of the real world of objects like spheres, tori, projective spaces or Möbius bands. For this reason we have avoided above the study of these axiomatic theories, replacing them by the simpler definition of manifolds as submanifolds of Euclidean spaces, considered up to suitable equivalence relation.

Similarly to the above story of the abstract manifolds, we will see that an "abstract group" is just a group of transformations.

2.1 Groups and groups of transformations

Consider a set X, for instance the set of points of a geometric object like a triangle, a cube, a sphere, etc.

Transformation. A *transformation* a of a set X is a bijection $a : X \to X$.

Examples. **a)** Every permutation of n distinct points is a transformation of that set of points.
b) The function $a(x) := x^n$ is a transformation of $\mathbb{R}$ if and only if n is odd.
c) Every symmetry of an equilateral triangle is a transformation of its points (by symmetry, we mean every rotation around its centre sending the triangle to itself and every reflection with respect to the triangle altitudes).
d) A linear map $A : \mathbb{R}^n \to \mathbb{R}^n$ is a transformation of the vector space $\mathbb{R}^n$ if and only if $\det A \neq 0$.

Since any transformation a is a bijection of X, its inverse transformation, a^{-1}, is always well defined: It sends each element of X to its preimage.

Product of Transformations. The *product* of two transformations a and b of a set X is the transformation $a \cdot b$ obtained if we apply to X the transformation b and then the transformation a: $(a \cdot b)(x) = a(b(x))$ for every $x \in X$.

Remark. It is not true, in general, that every pair a, b of transformations of X commutes. That is, it is not always true that the transformation $a \cdot b$, obtained as the product of the transformations a and b, coincides with the transformation $b \cdot a$, obtained as the product of b and a. For instance, for an equilateral triangle whose vertices are labelled by 1, 2 and 3 anti-clockwise, the transformations defined by the permutations

$$a = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$$

do not commute. The transformation a is a rotation by the angle $2\pi/3$ around the centre of the triangle in the anti-clockwise direction, and the transformation b is a reflection around the triangle altitude containing the vertex 3. The products of these transformations are

$$a \cdot b = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \qquad b \cdot a = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix},$$

so $a \cdot b \neq b \cdot a$.

Transformation groups. A collection G of transformations of X is called a *transformation group* if the product of any pair of transformations $a, b \in G$ belongs also to G, $a \cdot b \in G$, and if the inverse of each transformation $a \in G$ belongs also to G, $a^{-1} \in G$.

Examples. **a)** The set of all the transformations (permutations) of a set of n points is a transformation group usually called *the symmetric group* $S(n)$.
b) The set of all symmetries of an equilateral triangle is a group of transformations consisting of six symmetries and called the *symmetry group* of that triangle. Among its symmetries, the three anti-clockwise rotations around the centre by an angle $2k\pi/3$, for $k = 0, 1, 2$, form also a group of transformations. However, the set of symmetries of the equilateral triangle formed by the 3 reflections around its altitudes is not a group of transformations.
c) Similarly, the set of all symmetries of a geometric figure (like a tetrahedron, a sphere, an ellipsoid, etc.) is a group of transformations, called the *symmetry group* of that figure.

Abstract groups. Two groups of transformations (one of a first set of elements, the other of a second set, where even the number of elements of these sets may differ) are considered as the same "abstract group", if they are *isomorphic*, that is, if there is a bijective map of one group of transformations onto the other, sending the product of two transformations of the first group to the product of the images of these transformations in the second group.

Remark (On the axiomatic definition of group). Trying to generalise the naive groups of transformations, the algebraists created an axiomatic definition of "abstract groups". A *group* G is a non empty set G equipped with a binary operation that verifies:
– For all a, b in G, $a \cdot b$ is also in G;
– For all a, b, c in G, we have $(a \cdot b) \cdot c = a \cdot (b \cdot c)$;
– There is an element e in G, such that for all a in G, we have $e \cdot a = a \cdot e = a$;
– For each a in G, there exists an element b in G such that $a \cdot b = b \cdot a = e$.

Of course, these axioms imitate the evident properties of the transformations.

Cayley theorem is an elementary exercise for the axiom lovers, stating that the "axiomatic abstract groups" contain nothing new:

Cayley theorem. *Each (axiomatic) abstract group may be represented isomorphically by some group of transformations.*

Proof. It suffices to consider the action of the abstract group G on the set G, representing any element $g \in G$ as the left multiplication $L_g : G \to G$ sending every $h \in G$ to $L_g h = gh \in G$. □

Thus, the abstract groups of the axiomophiles are just the transformation groups considered up to isomorphisms. However, once the groups of transformations are introduced and understood, in some cases it may be useful to use also the axiomatic definition of group, as we will do sometimes.

2.2 Subgroups

A subset of a group G is said to be a *subgroup* of G, if itself is a group with respect to the same binary operation of G.

Example. The group of rotations of the equilateral triangle is a subgroup of the symmetry group of the equilateral triangle.

To study the properties of a group and to find its differences and relations with other groups, it is often useful to know and to study its subgroups.

Exercise. Prove that *a subset H of a group G is a subgroup of G if and only if the product of any two elements of H belongs to H, and for any element $a \in H$ its inverse a^{-1} belongs also to H.*

Order of a Subgroup. The *order* of a finite group G, denoted $|G|$, is the number of its elements. A non finite group has, by definition, infinite order.

Example. The symmetric group $S(n)$, formed by the permutations of n objects, has order $n!$.

When one studies finite groups, it is important to know their orders and the orders of their subgroups. For that, the following theorem is very useful:

Lagrange Theorem. *If H is a subgroup of a finite group G, then the order of H divides the order of G.*

Proof. The subgroup H defines an equivalence relation in G: Two elements g_1 and g_2 are equivalent if and only if there exists $h \in H$ such that $g_1 = g_2 \cdot h$. The equivalence classes of this equivalence relation are called *left cosets* and have the following description:

$$g \cdot H = \{g \cdot h : h \in H\} \ .$$

So, all the left cosets contain the same number of elements, namely the order of H, $|H|$, and hence the order of G is equal to the number of classes multiplied by $|H|$. Thus, $|H|$ divides $|G|$. $\square$

Example (The group of rotations of the cube). Let G be the group of rotations of the cube. A rotation is determined by the image of one first vertex (8 possibilities) and by the image of one of its contiguous vertices —whose image has to be contiguous to the image of the fist vertex— (3 possibilities). Hence, $|G| = 8 \cdot 3 = 24$.

So, the possible orders of the subgroups of G are 1, 2, 3, 4, 6, 8, 12 and 24. The subgroups of order 1 and 24 are the trivial subgroups $\{\mathrm{Id}\}$ and G.

Each subgroup of order 2 is formed by the identity and a rotation by π either around an axis through the centre of the cube and the centre of a face, or around an axis through the centre of the cube and the centre of an edge – Fig. 2.1. We can choose 3 axes in the first way, and 6 axes in the second one. So, there are 9 subgroups of order 2.

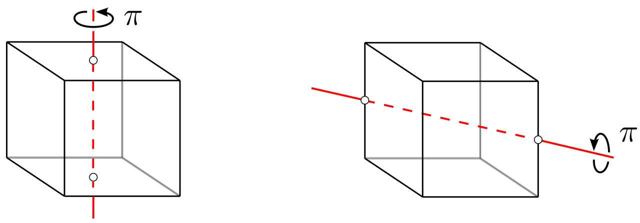


Figure 2.1: The subgroups of order 2 of the group of rotations of the cube.

The subgroups of order 3 consist of rotations by $2k\pi/3$ ($k = 0, 1, 2$) around the axes that join two opposite vertices – Fig. 2.2. Then there are 4 such subgroups.

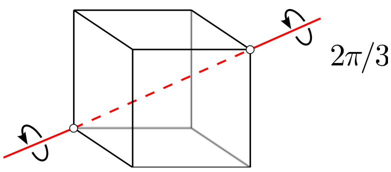


Figure 2.2: The subgroups of order 3 of the group of rotations of the cube.

The rotations by $\pi/2$ around an axis through the centre of the cube and the centre of a face generate a subgroup of order 4. Since there are three such axes, there are three such subgroups of order 4.

Exercise. There is still another subgroup of order 4 and a subgroup of order 12 in the group of rotations of the cube. Find them.

Remark. The fact that a natural number k divides the order of a group does not imply the existence of a subgroup of order k. For example, the group of rotations of the cube (of order 24) contains no subgroup of order 8.

2.3 Group Homomorphisms, Kernel and Image

Homomorphism. Let G and H be two groups whose respective identities are denoted e_H and e_G. A map $\varphi : G \to H$ is a *homomorphism* if it preserves the group structure:

$$\varphi(e_G) = e_H, \quad \text{and} \quad \varphi(a \cdot b) = \varphi(a) \cdot \varphi(b) \ , \ \forall a, b \in G \ .$$

Example. Given $\alpha \in \mathbb{R}$, the exponential function $\varphi : t \mapsto e^{\alpha t}$ is a homomorphism of the additive group $\mathbb{R}$ into the multiplicative group $\mathbb{R} \smallsetminus \{0\}$.

Example. Let G be the group of rotations of the cube. To every rotation of the cube there corresponds a permutation of its vertices. This define a group homomorphism $G \to S(8)$. Similarly, considering the permutations of the 6 faces of the cube, of the 4 main diagonals and of the 2 tetrahedrons inscribed in the cube, we get group homomorphisms from G to $S(6)$, $S(4)$ and $S(2)$.

Image. We recall that the *image* of a map $f : X \to Y$ is the subset of Y formed by all the possible values of f: $\operatorname{Im} f = \{f(x) : x \in X\}$.

Kernel. The *kernel* of a group homomorphism $\varphi : G \to H$ is the subset of G defined by: $\operatorname{Ker}\varphi := \{g \in G : \varphi(g) = e_H\}$.

Theorem. *The kernel* $\operatorname{Ker}\varphi$ *of a group homomorphism* $\varphi : G \to H$ *is a subgroup of* G.

Proof. For every $k, \ell \in \operatorname{Ker}\varphi$, we have $\varphi(k \cdot \ell) = \varphi(k) \cdot \varphi(\ell) = e_H \cdot e_H = e_H$, and hence $k \cdot \ell \in \operatorname{Ker}\varphi$. Moreover, for any $a \in \operatorname{Ker}\varphi$ we have

$$e_H = \varphi(e_G) = \varphi(aa^{-1}) = \varphi(a)\varphi(a^{-1}) = e_H\varphi(a^{-1}) = \varphi(a^{-1}),$$

proving that $a^{-1} \in \operatorname{Ker}\varphi$. □

Exercise. Prove that *the image* $\operatorname{Im}\varphi$ *of a group homomorphism* $\varphi : G \to H$ *is a subgroup of* H. Hint: The proof is very similar to preceding one.

Isomorphism. A homomorphism of groups $\varphi : G \to H$ is said to be an *isomorphism* if φ is invertible and its inverse φ^{-1} is a homomorphism.

Exercise. Show that the natural homomorphism (of the previous example) from the group of rotations of a cube to the symmetric group $S(4)$ is an isomorphism (proving that they represent the same abstract group).

An isomorphism of a group to itself is called *automorphism.* It is like a "symmetry" of the group, since the group structure is preserved.

Exercise. Fix an element g of a group G. Is the map $c_g : h \mapsto ghg^{-1}$ an automorphism of G? If yes, what is its inverse automorphism?

2.4 Group Actions and Flows

Group Actions. Let G be a group and M a set. We say that *a group G acts on a set M* if to each element g of G there corresponds a transformation $T_g : M \to M$ of the set M, in such a way that to the product of any two elements of the group corresponds the product of the transformations corresponding to these elements: $T_{fg} = T_f T_g$.

In other words, *the action of a group G on a set M* is a homomorphism of the group G into the group of all transformations of the set M.

Observe that a group action determines a map $T : G \times M \to M$ assigning to the pair (g, m) the point $T_g m$, which is denoted gm by short.

Orbits. Given a point m of M, the subset $\{gm : g \in G\}$ of M obtained from the action on m of all elements of the group G is called the *orbit of the point m* (for the given action), and is denoted Gm.

The points that are orbits are called *fixed points* of the action. A group action is said to be *transitive* if there is only one orbit (the whole set M).

Example. Given two non colinear vectors a, b of $\mathbb{R}^3$, assign to $(s, t) \in \mathbb{R}^2$ the transformation of $\mathbb{R}^3$ given by $T_{(s,t)}x = x + sa + tb$. This defines an action of the additive group $G = \mathbb{R}^2$ on the space $M = \mathbb{R}^3$. Each orbit is a plane containing the directions of a and b.

Problem. Find the orbits of the group of rotations of the plane about zero.

Problem. Let M be the set of matrices of linear maps from a vector space into itself. How does the group of linear coordinate changes act on M?

Answer. The coordinate change g sends the matrix m to $T_g m = gmg^{-1}$. The orbit of a matrix m is formed by all matrices that represent the same linear map (than m) in different linear coordinate systems.

Action by conjugation. The set formed by all the conjugates of an element g of a group G is called the *conjugacy class* of g. These elements have the same properties and behave in the same way as g in the group G. For example, in a group of transformations, the conjugates STS^{-1} of a given transformation T have the same geometric properties than T.

Thus the orbits of the action of a group G on itself by conjugation (taking all its *inner automorphisms* $c_g : h \mapsto ghg^{-1}$) are the conjugacy classes of G.

One-parameter groups. A *one-parameter group of transformations of a set M* is an action of the group of all real numbers on M.

Thus a one-parameter group of transformations of the set M is a collection $\{g^t\}$ of transformations $g^t : M \to M$, parametrised by the real parameter $t \in \mathbb{R}$, satisfying the *group property*: for any real numbers t and s

$$g^{t+s} = g^t g^s.$$

Example. The group property is obvious for the following examples.
1. $M = \mathbb{R}$, g^t is translation by $3t$ (i.e., $g^t x = x + 3t$);
2. $M = \mathbb{R}$, g^t is dilation by a factor e^t (i.e., $g^t x = e^t x$);
3. $M = \mathbb{R}^3$, g^t is translation by the vector ta ($a = (a_1, a_2, a_3)$). Thus g^t is the motion with constant velocity vector a (i.e., $g^t x = x + ta$).

Exercise. Find the orbits of the previous three one-parameter groups.

Phase flow. A one-parameter group of transformations of the set M is also called a *phase flow* with phase space M. The orbits of the phase flow are called *phase curves* (or *trajectories*).

Definition. A *phase flow of diffeomorphisms* is a one-parameter transformation group whose elements are diffeomorphisms satisfying the additional condition that $g^t x$ depends smoothly on both arguments t and x.
(Our diffeomorphisms will be often linear transformations.)

Examples. 1. $M = \mathbb{R}^2$, g^t is rotation about 0 by the angle t.
2. $M = \mathbb{R}^2$, g^t is scalar multiplication by e^{kt}: $g^t(x, y) = (e^{kt}x, e^{kt}y)$.

Problem. Let $\alpha, \beta \in \mathbb{R}$. Prove that the *quasi-homogeneous dilations* of the plane, $g^t(x, y) = (e^{\alpha t}x, e^{\beta t}y)$, form a phase flow of linear transformations. Describe the orbits for the cases $\alpha = 1, \beta = 2$ and $\alpha = 1, \beta = -1$.

Consider a one-parameter group $\{g^t\}$ of diffeomorphisms of a domain M.

Definition. The *phase velocity vector of the flow* $\{g^t\}$ at the point x in M is the velocity with which the point $g^t x$ leaves x, i.e.,

$$v_g(x) := \frac{d}{dt}|_{t=0}(g^t x)\,.$$

Phase velocity field. The phase velocity vectors of a flow at all points of the domain M form a smooth vector field (since $g^t x$ depends smoothly on t and x). It is called the *phase velocity field*.

EXERCISE. Find the velocity vector fields of the following flows on the line: $g^t x = x + 3t$, $g^t x = e^{-t}x$, $g^t x = e^{4t}x$ and $g^t x = 4e^t x$

ANSWER. $v(x) = 3$, $-x$, $4x$ and $4x$, but $g^t x = 4e^t x$ is not a flow (why?).

So *each phase flow of diffeomorphisms determines a differential equation*, defined by the phase velocity vector field, $\dot{x} = v_g(x)$. Its solutions are the motions of the points of M under the action of the phase flow:

Theorem. *Given a point x_0, the map $\varphi : \mathbb{R} \to M$ defined as $\varphi(t) = g^t x_0$ is a solution of the equation $\dot{x} = v_g(x)$ with initial condition $\varphi(0) = x_0$.*

Proof. It is a consequence of the group property:

$$\frac{d}{dt}\Big|_{t=r} g^t x \;=\; \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} g^{r+\varepsilon} x \;=\; \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} g^{\varepsilon}(g^r x) \;=\; v_g(g^r x). \qquad \square$$

For each smooth vector field v in a compact manifold M there is a flow whose velocity field is v. I is called the *flow of the vector field v*.

For general M a smooth vector field has not always a flow (see p. 231), though locally its flow always exists. Vector fields and their (local) flows are the key objects used to define the "Lie derivative" (see Chapter 7).

Solving differential equations by group actions. We shall illustrate the simplest case of solving a differential equation that is invariant under a flow of diffeomorphisms.

Example. A differential equation $\dot{x} = v(t, x)$ determines a field of directions in the plane whose slope at the point (t, x) is $v(t, x)$. If the direction field is invariant under the flow of the translations along the x-axis (i.e., if the slopes depend only of t) then the equation is just $\dot{x} = v(t)$. In this case the solution with initial condition $\varphi(t_0) = x_0$ is given by

$$\varphi(t) = x_0 + \int_{t_0}^{t} v(\tau)d\tau.$$

This example is a particular case of the following theorem (see [12]).

Theorem. *If a direction field in the plane is invariant under a flow of diffeomorphisms* (i.e., if each diffeomorphism of the flow sends the field to itself)*, then the differential equation defined by this direction field can be integrated explicitly in a neighbourhood of each non fixed point of the flow.*

This general case reduces to the above example by a "cleaver change of coordinates" in the plane (i.e., by a suitable diffeomorphism that locally sends the flow to a flow whose orbits are parallel lines). We refer the reader to [12, § 6] where the solution of homogeneous and quasi-homogeneous differential equations (whose direction fields are invariant under the respective groups of dilations $g^t(x, y) = (e^t x, e^t y)$ and quasi-homogeneous dilations $g^t(x, y) = (e^{\alpha t}x, e^{\beta t}y)$) is explained in a detailed way and including examples.

2.5 Relativity principle and invariance

As we mentioned above, two isomorphic groups represent the same abstract group. For instance, one neglects the difference between the group of the 6 symmetries of the equilateral triangle ABC in Paris and the group of the 6 symmetries of the equilateral triangle ЭЮЯ in Moscow – Fig 2.3.

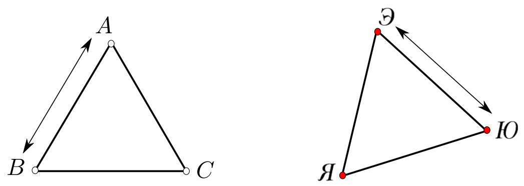


Figure 2.3: The symmetry groups of the equilateral triangles in Paris and in Moscow are isomorphic.

The relativity principle says that the scientific facts should not depend on notations, coordinate systems, units of measurement and other occasional circumstances.

A similar difference of ideas occurs in abstract and concrete mathematics at all levels. We recall some examples.

For the *affine spaces* (which may be defined by long lists of axioms, but are simply the vector spaces considered up to the change of the origin of the vectors) the Manhattan address *42 street - 5th avenue* is not a geometric description. It belongs to the coordinate system and therefore belongs more to the vector space structure introduced by the numbering of the streets in Manhattan rather than to its affine real world geometry.

Such coordinate systems are usually called "Cartesian", but they had been known and used before Descartes started his propaganda on their usefulness (for the fight of algebraists against geometry). Namely, Cartesian coordinates had been used by the Roman army in their camping to indicate the place of every legion, cohort and so on, in the crossings of their north-south and west-east directed streets. Some of them are still visible in the geography of the old Parisian streets in the Sorbonne area, the "Quartier Latin", reflecting the old Roman streets of the times of Cesar. The origin of this "Cartesian" coordinate system was located near the intersection of (the present streets) rue Saint Jacques and rue des Écoles. It is still fixed by the inscription "Jeux Descartes" (Cartesian games) referring not to the mathematician, but to the small shop "Jeux des cartes" (cards packs).

The intersection at one point of the three medians of a triangle is a fact of affine geometry. Even if we prove it by adding the vectors representing the vertices, this fact is independent of the choice of coordinates.

While algebraic computations may be simpler in a chosen coordinate system, the geometric point of view is to work "intrinsically", avoiding the auxiliary quantities depending on the coordinate system.

In mathematics this geometric character of the work is called "group invariance", while in physics the corresponding term is "relativistic invariance" (or "relativity principle" accordance). Thus the two formulae $6 + 7 = 13$ and $\mathrm{VI} + \mathrm{VII} = \mathrm{XIII}$ state the *same mathematical fact*, and the physical laws should be independent of the use of unities. Be the lengths measured in centimetres or in miles, the Newton law $F = md^2x/dt^2$ is the same fact, but the corresponding coordinates may have very different values.

The Russian poet Mayakovski had already described this idea at the time of the First World War. According to him, "the person who discovered that $2 + 2 = 4$ was a great mathematician, even if he discovered it by counting the cigarette parts. The person using today the same formula to count much greater objects —like the locomotives— is not a mathematician at all!". The calculation of the product $123 \cdot 456$ in the decimal system is not a mathematical activity, although we use it frequently and although relativists use coordinate systems.

One can more easily understand the *coordinate space* $\mathbb{R}^n = \{(x_1, \ldots, x_n)\}$ than the abstract idea of the n-dimensional *vector space* $\mathbb{R}^n$. The difference is only the will to preserve the geometric facts from the arbitrariness of the choice of coordinate systems (for instance, the choice between metres and yards for the measures of the lengths).

Similarly, the *Euclidean vector space* structure, being less restrictive than the coordinate space structure of the product of n real lines, is still anti-relativistic if we are studying the affine geometry because such notions as angles between lines and lengths of segments are missing in affine geometry. Thus an affine transformation, like $g : (x, y) \mapsto (2x, y)$, may modify the angles and the lengths, preserving all the affine geometry (Fig. 2.4).

Fixing a particular structure on a manifold or space M, we choose a special geometry – it is the study of the properties (of the objects of M) which are invariant under the action of the group of transformations that preserve the given structure on M. Thus, for the same set $M = \mathbb{R}^n$ we can study the Euclidean vector space geometry or the real vector space geometry, the Euclidean affine space geometry, or the real affine space geometry. These

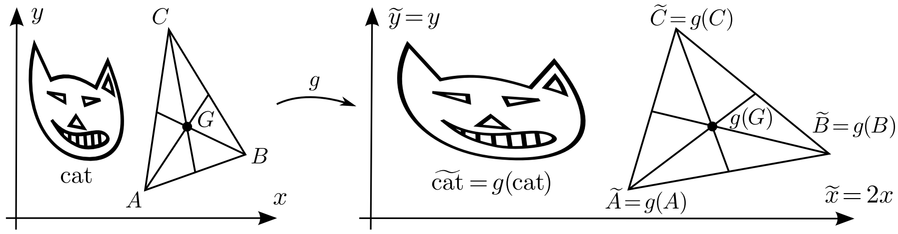


Figure 2.4: The action of an affine transformation g on different geometric objects, like on the centre of mass G of a triangle ABC.

different geometries are preserved by different transformations that, in each case, form a special group.

In the preceding four cases we get four geometric groups acting on $\mathbb{R}^n$:

$$\begin{array}{ccc} G_{\text{Euclidean,vector}} & \subset & G_{\text{real,vector}} \\ \cap & & \cap \\ G_{\text{Euclidean,affine}} & \subset & G_{\text{real,affine}} \end{array}$$

2.6 Invariant (normal) subgroups

A subgroup of a group G has relativistically good behaviour (being independent of the choice of the "coordinate system" and of any other occasional non-intrinsic choice of the type of Manhattan's description) if it remains the same under the "symmetries of G", that is, under the action of the group G on itself by inner automorphisms. Such a subgroup is said to be *normal* or *invariant*. The formal definition is:

Normal subgroup. A subgroup H of a group G is said to be *invariant* or *normal* if $gHg^{-1} = H$ for any $g \in G$.

It may also be formulated as $gH = Hg$, that is, for any $h \in H$ $\exists \widetilde{h} \in H$: $gh = \widetilde{h}g$. Hence, the left class $gH = \{gh : h \in H\}$ equals the right class Hg.

Example. The set G_A of symmetries of the cube that fix its vertex A is a subgroup of 6 elements of the group of 48 symmetries of the cube.
The subgroup G_A is not invariant: Take a symmetry g sending A to a different vertex $\widetilde{A}$, $gA = \widetilde{A}$. The conjugation by g does not fix the subgroup G_A, it sends G_A to the subgroup $G_{\widetilde{A}}$. Of course $G_{\widetilde{A}}$ is isomorphic to G_A, representing the same abstract group S(3), but it defines a subset of 6 symmetries of

the cube that is *different* from the subset defined by G_A. In Fig. 2.5, we illustrate the result of the conjugation by a symmetry g on a rotation $h \in G_A$, $\widetilde{h} = ghg^{-1}$.

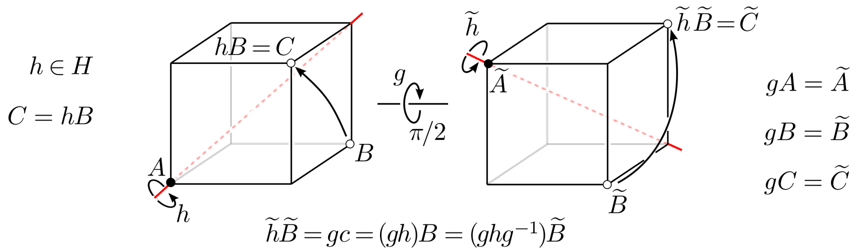


Figure 2.5: The action of a symmetry g on the subgroups: If g sends A to the vertex $\widetilde{A}$, the conjugation by g sends the rotation $h \in G_A$ to $ghg^{-1} = \widetilde{h} \in G_{\widetilde{A}}$.

Example. Each subgroup of order 3 of the group of rotations of the cube corresponds to one diagonal of the cube. These subgroups are not invariant, since they depend on the choice of the diagonal.

Exercise. Let G be the group of symmetries of the tetrahedron: $|G| = 24$. For each vertex of the tetrahedron there is a subgroup of G formed by those symmetries preserving that vertex. So, we get 4 subgroups of G, each one of order 6. Question: Are these subgroups invariant?

Exercise. Consider the symmetry group of the equilateral triangle. Besides the trivial subgroups, it has 3 subgroups of order 2 and one subgroup of order 3 (the group of rotations of the triangle). Question: Is any of these subgroups invariant?

In order to find invariant subgroups, the following theorem is useful.

Theorem. *The kernel of a homomorphism of groups is a normal subgroup.*

Proof. Let G and H be two groups, with respective unities e_G and e_H, and let $\varphi : G \to H$ be a group homomorphism. From the very definition of group homomorphism, we have that for every $g \in G$ and $h \in \mathrm{Ker}\,(\varphi)$:

$$\varphi(g \cdot h \cdot g^{-1}) = \varphi(g) \cdot \varphi(h) \cdot \varphi(g)^{-1} = \varphi(g) \cdot e_H \cdot \varphi(g)^{-1} = e_H \ .$$

This proves the invariance of $\mathrm{Ker}\,(\varphi)$: $g \cdot \mathrm{Ker}\,(\varphi) \cdot g^{-1} = \mathrm{Ker}\,(\varphi)$. $\square$

The converse holds:

Theorem. *Any normal subgroup H in G is the kernel of a homomorphism of G to some group* (this group, called the *quotient group G/H*, is the result of some kind of division of G by H, whence the notation G/H).

Proof. The construction is very simple (see Fig. 2.6). The elements of the quotient group G/H are the sets $gH = \{gh : h \in H\}$, for arbitrary fixed elements g of G. The group G is thus subdivided into disjoint subsets gH. In fact, if g_1H intersects g_2H, then $g_1H = g_2H$ because for a common element $k = g_1h_1 = g_2h_2$, we have $g_1 = kh_1^{-1}$, $g_2 = kh_2^{-1}$, and hence $g_1H = kH = g_2H$. To complete the proof, we need the following proposition.

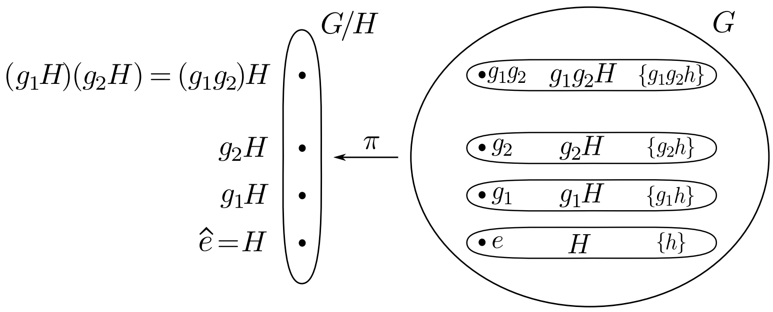


Figure 2.6: The "factorisation homomorphism" $\pi : G \to G/H$.

Quotient group. The set whose elements are the above disjoint subsets is called *the quotient group* and is denoted G/H.

Proposition. *The quotient group G/H possesses a natural group structure.*

Proof. The product of two elements of G/H is defined as the product of their representatives:

$$(g_1H)(g_2H) = (g_1g_2)H.$$

Indeed, since $g_1H = Hg_1$, we have that for any $h_1 \in H$ there exists $\widetilde{h}_1 \in H$ such that $g_1h_1 = \widetilde{h}_1g_1$ and we get therefore

$$(g_1h_1)(g_2h_2) = (\widetilde{h}_1g_1)(g_2h_2) = \widetilde{h}_1(g_1g_2)h_2 = (g_1g_2)\widetilde{\widetilde{h}}_1h_2,$$

for a suitable element $\widetilde{\widetilde{h}}_1 \in H$. This proves that the product is independent of the choices of the representatives. Clearly, the identity element is $\widehat{e} = H$. $\square$

Our construction provides the natural homomorphism $\pi : G \to G/H$ onto the quotient group. The kernel of this homomorphism is the invariant subgroup H. The theorem is proved. $\square$

Problem. Which of the four geometric groups of transformations of $\mathbb{R}^n$ (described in p. 41) are normal (relativistic) subgroups ?

Solution. The answers are all negative: The same affine space can be made a vector space by different choices of the origin, and those different, but isomorphic, subgroups H preserving the vector spaces are permuted by the “coordinate change” g (acting as gHg^{-1}), while they ought to be invariant under this conjugation, if the subgroup H was normal. Hence the subgroup of linear transformations is not a normal subgroup of the group of affine transformations, be it for the general case or for the Euclidean one.

Similarly, different choices of the Euclidean metric (providing isometric vector spaces or affine spaces) are not relativistically invariant under the non Euclidean (linear or affine) transformations, but are permuted by them.

The subgroup H of the Euclidean linear or Euclidean affine transformations of the relevant vector space or affine space, respectively, is hence not preserved by the general (linear or affine, respectively) isomorphisms of the non Euclidean object, but are rather permuted by it.

Hence the Euclidean linear (respectively affine) group of transformations is not a normal subgroup of the corresponding general linear (respectively affine) group of transformations.

Remark. A different interesting subgroup of the affine group is the commutative subgroup of translations H, acting on $\mathbb{R}^n$ as $h(x) = h + x$, for the action of the element $h \in H$.

The subgroup $H \approx \mathbb{R}^n$ preserves the Euclidean structure of the space $\mathbb{R}^n = \{x\}$, being thus a part of the Euclidean affine group of transformations.

Problem. Is H a normal subgroup of the group of affine transformations of $\mathbb{R}^n$? And of the group of its Euclidean affine transformations?

Hint. Associate to an affine transformation $g(x) = Ax + b$ in $\mathbb{R}^n$ its linear part $A : \mathbb{R}^n \to \mathbb{R}^n$ and show that the subgroup H is the kernel of this homomorphism, being therefore a normal (invariantly defined) subgroup.

2.7 Invariant subgroups of the Platonic polyhedra symmetry groups

There are only 5 regular convex polyhedra in Euclidean space $\mathbb{R}^3$: The *tetrahedron* (with 4 faces), the *octahedron* (with 8 faces), the *cube* (with 6 faces), the *icosahedron* (20 faces) and the *dodecahedron* (12 faces) – Fig. 2.7.

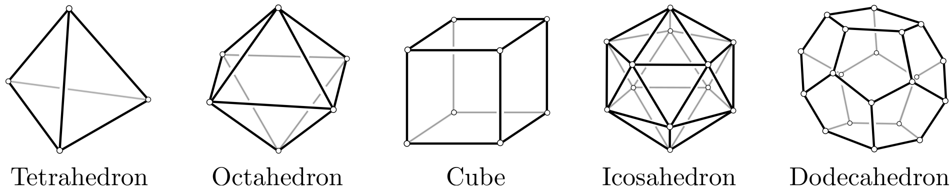


Figure 2.7: Platonic polyhedra.

The following table lists the numbers of vertices (v), faces (f) and edges (e) of the Platonic Polyhedra.

	tetrahedron	**octahedron**	**cube**	**icosahedron**	**dodecahedron**
v	4	6	8	12	20
e	6	12	12	30	30
f	4	8	6	20	12

Remark. The *Euler relation*, $v-e+f=2$, holds for the numbers of vertices, edges and faces of the Platonic Polyhedra. In fact it is true for any convex polyhedron (homeomorphic to a sphere). The reader should verify that for each polyhedron of Fig. 2.8 (homeomorphic to a torus) the quantity $v-e+f$ has the same value ($\neq 2$).

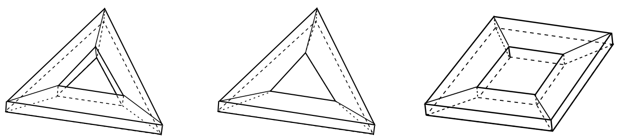


Figure 2.8: Three polyhedra homeomorphic to a torus.

In Fig. 2.9, the number of faces of a polyhedron is increased by subdividing one of its faces. The value of $v-e+f$ remains constant even when some new faces are non-convex, provided that the interior of each new face is homeomorphic to an open disc.
The value of $\chi := v-e+f$, called *Euler characteristic* of the polyhedron, depends only on the topological nature of the surface rather than on its polyhedral realisation.

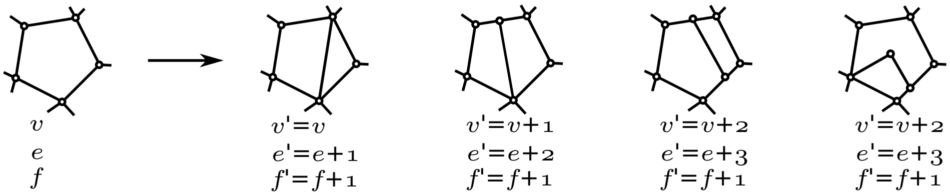


Figure 2.9: Subdivision of a face of a polyhedron into smaller polygonal faces.

It is a very important problem to find all the normal subgroups of the symmetry groups of the five regular polyhedra.

Octahedron-Cube. The geometry provides, with no calculation at all, the following evident but nontrivial information:

Proposition. *The normal subgroups of the groups of symmetries of the cube and of the octahedron are isomorphic.*

Proof. The 6 centres of the 6 faces of a cube form the 6 vertices of an inscribed octahedron (similarly, the centres of the 8 faces of the octahedron form the 8 vertices of an inscribed cube) – Fig. 2.10. The cube symmetries induce therefore the symmetries of the inscribed octahedron, and vice-versa. Thus the symmetry groups of the cube and of the octahedron are isomorphic. Hence they have isomorphic normal subgroups. □

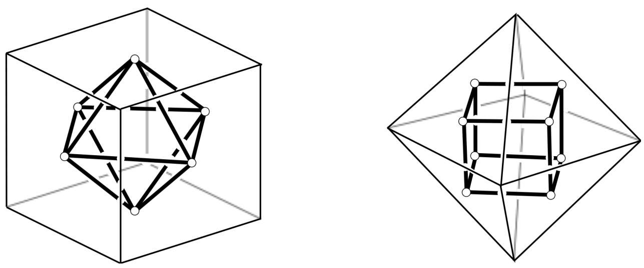


Figure 2.10: The inscriptions of the octahedron into the cube and of the cube into the octahedron.

Normal subgroups of the symmetry group of the cube

One of the normal subgroups of the symmetry group of the cube (or of the octahedron) is the group of rotations, containing 24 of the 48 symmetries.

To find another normal subgroup (whose existence is not evident), we shall inscribe two tetrahedra in the cube: Their vertices are 4 vertices of the cube that are pairwise joint by 6 diagonals of its 6 faces – Fig. 2.11.

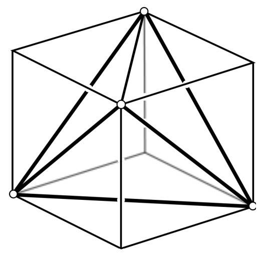


Figure 2.11: One of the two tetrahedra inscribed in the cube. To master this object, one should try to calculate the volume of the intersection of this tetrahedron with the second tetrahedron inscribed in the same cube of edge length 1.

The symmetries of the cube act on this set of two tetrahedra by permuting them. The group of permutations of two tetrahedra is the 2-elements group $\mathrm{S}(2) = \mathbb{Z}_2$. We have thus constructed a homomorphism

$$G \longrightarrow \mathbb{Z}_2$$

of the group of 48 symmetries of the cube onto the group of permutations of two tetrahedra. Its kernel is the subgroup of 24 cube symmetries that send the first tetrahedron to itself (and therefore send the second also to itself).

This subgroup is normal (being the kernel of a homomorphism of groups) and is different from the group of rotations (since some rotations permute the two inscribed tetrahedra).

It is interesting to study further the intersection of the two normal subgroups of 24 elements of the 48-elements group of symmetries of the cube.

The four long diagonals of the cube are permuted arbitrarily by the 24 rotations of the cube. Among these permutations, there are 12 even permutations that preserve each tetrahedron and form a new interesting invariant subgroup: The intersection of two invariant subgroups is automatically invariant, since the action on each of those subgroups preserves each of them, and hence the action on their intersection sends it to the intersection of their images, that is, sends it to itself.

We leave to the reader the interesting study of the normal subgroups of the group of 12 even permutations of 4 elements to which we have arrived above,

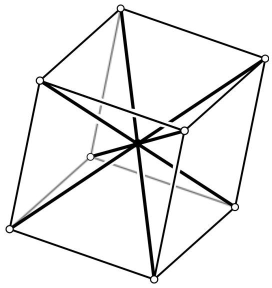


Figure 2.12: The four long diagonals of a cube.

and that of the remaining normal subgroups of the group of symmetries of the tetrahedron and of the cube (and hence of the octahedron).

Icosahedron-Dodecahedron.* The group G formed by the 120 symmetries of the dodecahedron is isomorphic to the group of symmetries of the inscribed icosahedron whose 12 vertices are the centres of the 12 pentagonal faces of the dodecahedron. Similarly, the centres of the 20 triangular faces of an icosahedron are the 20 vertices of an inscribed dodecahedron.

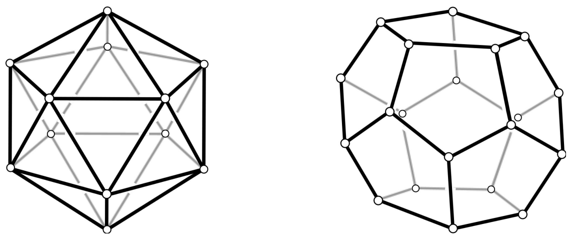


Figure 2.13: The icosahedron (20 triangular faces, 12 vertices and 30 edges) and the dodecahedron (12 regular pentagonal faces, 20 vertices and 30 edges).

In this symmetry group G of 120 elements, the subgroup of 60 rotations is a natural normal subgroup. Its invariance follows from the natural action of the 120 symmetries of G on the two orientations of the ambient space $\mathbb{R}^3$: It provides the orientation reversing homomorphism $G \longrightarrow \mathbb{Z}_2$, whose

*The names come from the Greek "icos"= 20 and "dodeca"= 12, counting their faces. The regular disposition of the vertices of the icosahedron and of the dodecahedron on their circumscribed spheres is used for the construction of nuclear bombs (to send many parts of the critical mass together simultaneously). The well known football use of this geometry was a later discovery, inspired by these military applications.

kernel is the subgroup G^+, consisting of the 60 rotational symmetries of the dodecahedron.

We have thus constructed a normal subgroup G^+ of 60 elements in G.

Claim. *There is no other nontrivial normal subgroup in this case* (the two trivial subgroups —the group itself and the unity element— are always normal subgroups in any group).

This remarkable fact is the core point in the topological proof of the celebrated Abel's theorem on the unsolvability of the algebraic equations of degree 5 in terms of radicals and rational operations (see Chapter 13).

To prove this fact, it is useful to consider the 5 Kepler cubes inscribed in the dodecahedron. To construct a Kepler cube, one takes 8 of the 20 vertices of the dodecahedron and connect them by 12 diagonals of the dodecahedron faces (one diagonal for face) in the following way: The 3 diagonals of the 3 faces meeting at a vertex V of the dodecahedron are obtained from one of them by the rotations of the dodecahedron, of 120°, that preserve V. In this way, we start from an end-point V of an arbitrary diagonal of one face of the dodecahedron, and then we construct the 2 neighbouring diagonals on the 2 other faces containing V. We have thus the 3 diagonals starting at V.

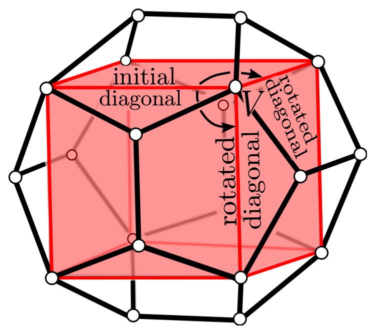


Figure 2.14: One of the Kepler cubes inscribed into the dodecahedron.

Following each of these 3 diagonals to their respective opposite end-points, we construct 2 new diagonals of 2 new faces at each end-point, obtaining 6 new diagonals whose opposite end-points coincide by pairs at 3 vertices from which we construct the 3 remaining diagonals. These 12 diagonals are the edges of our Kepler cube.

According to the five possible choices of the initial diagonal at the first face, we get the five Kepler cubes. *

The dodecahedron symmetries permute these five cubes. There are 120 symmetries and 120 permutations. *Besides the identity, no symmetry preserves the five Kepler cubes.* The homomorphism of the dodecahedron symmetry group to the group of permutations of the 5 cubes is therefore an isomorphism: Any permutation of the five Kepler cubes can be realised by a convenient symmetry of the dodecahedron.

For instance, the 60 rotational symmetries are sent to the subgroup of the 60 even permutations of the 5 cubes. Hence, the group of rotational symmetries of the dodecahedron is isomorphic to the group the 60 even permutations of five elements. The isomorphism is provided by the permutations of the 5 Kepler cubes.

The even permutations form a normal subgroup, because they form the kernel of the homomorphism to the orientation reversal group $\mathbb{Z}_2$. The claim of Abel is that there exists no other normal subgroup in the group of symmetries of the dodecahedron, and hence in the groups $S(5)$ and $S^+(5)$ of arbitrary and of even permutations of 5 elements, respectively.

This geometric statement has an easy geometric proof. Indeed, if a normal subgroup H was containing some symmetry h, it would also contain all its conjugated symmetries ghg^{-1}, whatever be the symmetry g.

Thus if the normal subgroup H contains, say, an element preserving an edge of the dodecahedron, but exchanging its end-points, then this subgroup should contain all such elements associated to all the 30 edges of the dodecahedron, since g can be a symmetry sending the original edge to any other. In this case, the normal subgroup H should contain the products of all such elements, and then at least 60 elements. If H contains no orientation reversing element, then H is the group of rotations (otherwise H coincides with the whole symmetry group of the dodecahedron).

Repeating this reasoning with other elements h (say, with the rotation around one vertex, whose conjugates are the similar rotations around all other vertices; or with the rotation of a face, conjugated to similar rotations of all

*Conversely, the dodecahedron can be obtained as "a house constructed from the cube" by adding the "roofs" (consisting each of two pentagonal plates) over every face of the cube. Kepler used the length relations between the inscribed regular polyhedra elements to obtain a model for the distribution of the diameters of the planetary orbits in the solar system. The 4th law of Kepler suggested one more planet between Mars and Jupiter, where one observes today a lot of asteroid orbits.

the other faces) we get immediately that any nontrivial normal subgroup of the subgroup of rotations is the whole subgroup of rotations, while any normal subgroup containing an orientation reversing element is the whole group of the 120 symmetries of the dodecahedron.

Remark. The geometrical content of Abel's theorem is the absence of any natural construction of a set of geometric objects, independent of the Kepler cubes, on which acts the group of symmetries of the dodecahedron. Of course, we could perform some construction of symmetric geometric objects at each of the Kepler cubes, but the resulting objects would be permuted by the dodecahedron symmetries.

Exercises. 1. Determine the quotient groups of the group of rotations of the cube by its normal subgroups.
2. Let O(3) be the orthogonal group of Euclidean space $\mathbb{R}^3$, SO(3) be its subgroup formed by the matrices of determinant 1. Is SO(3) an invariant subgroup of O(3)? If yes, determine the quotient O(3)/SO(3).
3. Once we fix a plane in Euclidean space $\mathbb{R}^3$, the group $SO(2) \simeq \mathbb{S}^1$ may be considered as a subgroup of SO(3) (the rotations preserving that fixed plane). Is SO(2) an invariant subgroup of SO(3)?

2.8 Abstract and "Naive" Definitions

H. Poincaré had already discussed the difficulties of the abstract mathematical notions for the beginners, as the simple fractions. He claimed that there exist only *two* methods to teach fractions: One should subdivide with a knife either an apple or a round pie – Fig. 2.15. According to Poincaré, all other "axiomatic" ways (like the Grothendieck ring or the Dedekind equivalence classes of pairs of integers) would naturally lead children to a simpler system of axioms where, for instance, $1/2 + 1/3 = 2/5$, as a large part of U.S. students think nowadays.

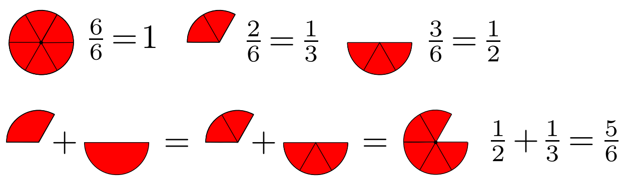


Figure 2.15: The Poincaré arithmetics of fractions.

Californian (anti-federal) law requires the knowledge of this arithmetics for the high-school students coming to the university: They ought to be able to divide 111 by 3 with no computer help. Washington senators rejected this anti-constitutional Californian requirement, as well as the teaching of any theory unknown to them, including the "racist" theory of the 3 phase states of the water (the federal curriculum including only the liquid and the solid states, "transformed one to the other in the refrigerators").

Investigating the stability of a dynamical system during a written exam, a student of Mathematics (in the 4th year at Paris 9 University), calculated the asymptotic expansion of solutions of complicated differential equations (using correctly the theories that Arnold taught him) to obtain the constant $4/7$. His subsequent results were, however, erroneous because he did not know whether $4/7$ was less than or greater than 1 (influencing the convergence of an integral involved in the study of asymptotics of a solution).

These stories explain the reason of the inclusion of elementary things on the preceding pages: We should not expect these elementary parts of mathematics to be studied earlier, since the present day educational system neglects to teach many general basic foundations of our science.

One more example of a difficult abstract mathematical notion is that of algorithm.

The corresponding naive object is the so-called Turing machine, which is a formalised description of a finite automate acting locally on a (possibly infinite) sequence formed with the symbols 0 and 1, deciding automatically at each step whether the next step will be performed one place to the left or one place to the right with respect the present one.

Initially, the sequence is filled by a finite number of symbols, called "the program". Its final situation is called "the answer".

There are hundreds of other more abstract definitions (and 100^2 theorems proving their equivalences). The main result of this algorithm theory is the following meta-mathematical statement, which, being a definition, cannot be proved: "All algorithms" are performed by one "universal" Turing machine. For that machine it was proved, at least, that it can do the works of any other Turing machine.

The difficulty is that the words "all algorithms" has here no other definition. It is a kind of belief that no other objects would be ever called by the word "algorithm".

Our choice in this book, to replace the abstract manifolds and groups by their original naive versions, is quite similar to the choice in algorithm theory to restrict the study to the Turing machine case. It is interesting that Turing himself insisted that the formal logic of an axiomatic theory is insufficient for the true scientific discoveries which, in practice, use a lot of unproved arguments taken from experiences.

2.9 Free Groups and Defining Relations

There are lots of situations in which no explicit description of the involved groups is known and one has no procedure to find such description. In many such cases one can represent implicitly such groups in an abstract way (without making use of particular properties of its elements) by means of formal sequences of symbols (words) and imposing relations between those sequences. For that the following notions are needed.

Free Groups. We start with a finite set of symbols $x_1, x_1^{-1}, \ldots, x_k, x_k^{-1}$, calling it *alphabet* (all results hold if it is infinite), where to x_j corresponds the symbol x_j^{-1}. A *word* is an ordered sequence of symbols of this alphabet (for example, $x_4x_3x_1x_4^{-1}x_3^{-1}$). The following "orthographic" rules are used.

The empty word (with no symbol) is allowed and denoted by e (or 1). A word is *reduced* if no symbol x_j stands adjacent to its associated symbol x_j^{-1}. For example, the words $x_4x_4x_1x_4^{-1}$ and $x_4^{-1}x_3^{-1}x_4x_3x_3$ are reduced, bot not $x_3x_4x_4^{-1}x_3$. The set of all reduced words that can be formed with symbols of our alphabet is turned into a group as follows.

The *product* of two words w_1, w_2 is the word obtained by setting w_2 after w_1, and suppressing all adjacent pairs x_j, x_j^{-1} till get a reduced word. It is clear that the *unit element* is the empty word and the *inverse* of a word is obtained from it by putting its symbols in the opposite order, replacing x_j with x_j^{-1} and x_j^{-1} with x_j. The associativity proof is left to the reader.

Definition. The obtained group is the *free group* of rank k (on k generators) F_k. The elements $x_1, \ldots, x_k$ form a *free system of generators* (any word is the product of the x_j and their inverses x_j^{-1}). Apart from the existence of inverses, $x_jx_j^{-1} = e$, no other relation exists between the generators.

To save paper one replaces the products $x_jx_j \cdots x_j$ (k times) with x_j^k and $x_j^{-1}x_j^{-1} \cdots x_j^{-1}$ (k times) with x_j^{-k}, writing $x^0 = e$.

Example. The free group on one generator x is the infinite cyclic group. It is isomorphic to the additive group $\mathbb{Z}$ (one can take the isomorphism $x^k \mapsto k$).

Every free group on more than one generator is non commutative (indeed, x_ix_j and x_jx_i are different words if $i \neq j$). Moreover, *all elements of a free group, except the unit element, have infinite order.*

Theorem. *Every group is isomorphic to a quotient group of a free group.*

Proof. Given a group G with a system of generators $g_1, \ldots, g_k$, take a free system of generators $x_1, \ldots, x_k$ of the free group F_k. The map φ that carries x_j to g_j and, in general $x_{j_1}^{s_1}x_{j_2}^{s_2} \cdots x_{j_\ell}^{s_\ell}$ to the element $g_{j_1}^{s_1}g_{j_2}^{s_2} \cdots g_{j_\ell}^{s_\ell}$ of G (where $s_i = \pm 1$) is a homomorphism of F_k onto G. Hence $G \simeq F_k/\mathrm{Ker}\,\varphi$. $\square$

Defining Relations. Since the normal subgroup $N = \operatorname{Ker}\varphi$ consist of the words whose image in G is the unit element, to the word $x_{j_1}^{s_1} x_{j_2}^{s_2} \cdots x_{j_\ell}^{s_\ell}$ in N corresponds the *relation* $g_{j_1}^{s_1} g_{j_2}^{s_2} \cdots g_{j_\ell}^{s_\ell} = e$ between the generators of G.

A *system of defining relations* of $G\,(\simeq F_k/N)$ is a collection of relations $g_{j_1}^{s_1} g_{j_2}^{s_2} \cdots g_{j_\ell}^{s_\ell} = e$ in G whose corresponding words $x_{j_1}^{s_1} x_{j_2}^{s_2} \cdots x_{j_\ell}^{s_\ell}$, together with their conjugates, generate the normal subgroup N. *Any other relation that links the generators of* G *is consequence of the defining relations.*

Example. Starting from the free group of one generator, say, x, the cyclic group of order k ($\simeq \mathbb{Z}/k\mathbb{Z}$) is given by the single defining relation $x^k = e$.

Example. The group of symmetries of an equilateral triangle is given by two generators ρ (a rotation) and σ (a reflection) with the defining relations $\rho^3 = 1$, $\sigma^2 = 1$ and $\rho\sigma\rho\sigma = 1$. This group is also given by two reflections σ_1, σ_2 with the defining relations $\sigma_1^2 = 1$, $\sigma_2^2 = 1$ and $(\sigma_1\sigma_2)^3 = 1$.

Remark. Specifying defining relations uniquely determines the normal subgroup N, and hence G. So one often says that "a group is given by a set of defining relations" or "is obtained from a group modulo a set of relations". This is a method of presenting a group "abstractly", that is, besides the group itself it provides all the groups isomorphic to it (to the pleasure of algebraists). However, in most cases one has very few information about a group given by defining relations. For example, every group can be given by defining relations in many distinct ways, but knowing whether two groups given by defining relations are isomorphic (or not) is an extremely hard problem. Moreover, if a group is presented by defining relations, we cannot determine, as a rule, if it is commutative or not, finite or infinite, and so on.

Chapter 3

Homotopy groups

In order to work with manifolds it is convenient to have "invariants" that permit to distinguish them and to understand better their most basic properties. Examples of such "invariants" are the "homotopy groups" of a manifold. These groups depend only on the "topological type" of the manifold (but not on its particular realisation).

Besides the construction of homotopy groups, we present the notions of "homotopy" and "homotopy equivalence", as well as some techniques and tricks to compute homotopy groups, applying them concretely in several important examples of manifolds.

3.1 Homotopy groups and fundamental group

We start from the k-dimensional cube

$$I^k = \{t \in \mathbb{R}^k : t = (t_1, \dots, t_k), \quad 0 \leqslant t_j \leqslant 1, \quad j = 1, 2, \dots, k\}.$$

Spheroid. A *k-dimensional spheroid* at a point $*$ of M is a continuous map

$$\varphi : I^k \to M$$

sending the whole boundary of the cube to the prescribed point: $\varphi(\partial I^k) = *$.

The *kth homotopy group* $\pi_k(M, *)$ has as its elements the homotopy classes of the k-dimensional spheroids at the point $*$ of M, that we introduce now.

Spheroid Homotopy. A *homotopy* between two spheroids, φ_0 and φ_1, is a continuous family of spheroids φ_s, for $0 \leqslant s \leqslant 1$, starting at φ_0 and ending at φ_1. So, it is a continuous map of the cube of dimension $k+1$ to M,

$$H : (I^{k+1} = I^k \times I) \to M, \quad H(t, s) \in M, \ t \in I^k, \ s \in I,$$

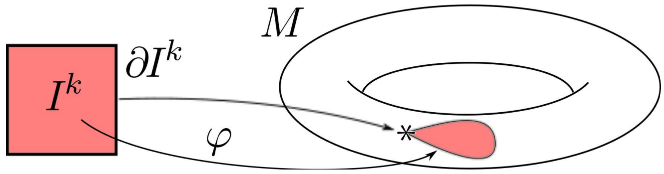


Figure 3.1: A 2-dimensional spheroid φ at the point $*$ of the torus M.

sending the whole boundary of the cube I^k to the prescribed point $*$ for any value of $s \in I$ ($H(\partial I^k \times I) = *$) and coinciding with φ_0 for $s = 0$ and with φ_1 for $s = 1$. In such case we say that the spheroids φ_0 and φ_1 are *homotopic*.

For $k = 1$, we get just the definition of homotopy between two loops.

Homotopy Classes. It is clear that the homotopy between spheroids is a (symmetric) equivalence relation. Its equivalence classes are called *homotopy classes* of the k-dimensional spheroids of M at its point $*$.

Homotopy Group. To define the group operation on these classes, we do the same thing as for the fundamental group operation – Fig. 3.2.

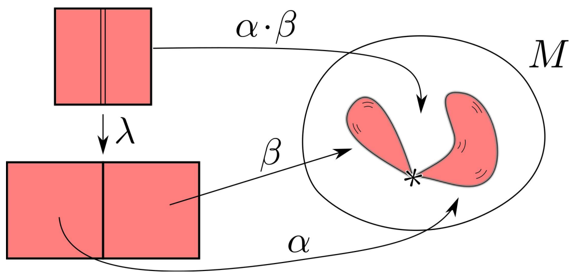


Figure 3.2: Multiplication of the spheroids α and β.

Consider two spheroids α and β as the maps of two neighbouring cubes in $\mathbb{R}^k$ that have a common face, and send the standard cube I^k to the union of the above two cubes in $\mathbb{R}^k$ by a standard linear map λ (transforming only the single coordinate t_1 which distinguishes the neighbouring cubes).

The resulting map of the standard cube I^k to M (produced by α at the left half-cube $0 \leqslant t_1 \leqslant 1/2$, and by β at the right half-cube) sends to the prescribed point $*$ of M all the boundary of the cube I^k, as well as its section $t_1 = 1/2$. It is therefore a spheroid of M at the point $*$.

The homotopy class of this new spheroid $\alpha \cdot \beta$ is independent of the particularities of the choices of the spheroids α and β in their homotopy classes. We get therefore an operation which associates to the two homotopy classes of spheroids of M at the point $*$, $[\alpha]$ and $[\beta]$, a new homotopy class $[\alpha \cdot \beta]$, called their product, $[\alpha] \cdot [\beta]$.

This is a group operation. The associativity $([\alpha] \cdot [\beta]) \cdot [\gamma] = [\alpha] \cdot ([\beta] \cdot [\gamma])$ is evident. The trivial element $[*]$ is the class of the trivial spheroid, which sends the whole cube to the prescribed point $*$:

$$[\alpha] \cdot [*] = [*] \cdot [\alpha] = [\alpha].$$

The inverse class $[\alpha]^{-1}$ is produced from $[\alpha]$ by the cube reflection sending t_1 to $1 - t_1$:

$$[\alpha]^{-1} \cdot [\alpha] = [\alpha] \cdot [\alpha]^{-1} = [*].$$

This homotopy can be constructed explicitly by replacing at each moment $0 \leqslant s \leqslant 1$ the map $\alpha \cdot \alpha^{-1}$ of the middle part $|t_1 - 1/2| \leqslant s/2$ of each segment of direction t_1 in the cube by its values at the points where $t_1 = 1/2 \pm s/2$.

We have thus defined the *homotopy group* $\pi_k(M, *)$.

Fundamental group. Case $k = 1$: The group $\pi_1(M, *)$ it is called the *fundamental group* of M. It describes the homotopy classes of the loops in M, that is, of the maps of the circle $\mathbb{S}^1$ to M, defined as continuous maps $\varphi : I \to M$ of the segment $I = [0, 1]$ to M that send the boundary $\partial I = \{0, 1\}$ to the prescribed point $*$ of M: $\varphi(\partial I) = *$.

We shall prove below (p. 76) that there are manifolds whose fundamental group is non commutative (for example, any compact surface of genus $g \geqslant 2$). However, there is an important difference with the homotopy groups $\pi_{k \geqslant 2}$:

Theorem 1. *For $k \geqslant 2$ the homotopy group $\pi_k(M, *)$ is commutative: The spheroids $\alpha \cdot \beta$ and $\beta \cdot \alpha$ are homotopic.*

Proof. The homotopy is explicitly shown in Fig. 3.3. $\square$

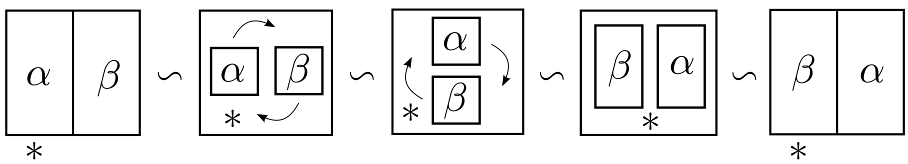


Figure 3.3: Proof of the commutativity of the homotopy groups.

Remark. This astonishing theorem is a manifestation of the strange general topologic phenomenon discussed at the end of this chapter: Things become simpler in higher-dimensional cases, “due to the presence of more space for the constructions”.

Since the homotopy groups π_k are commutative for $k \geqslant 2$, one uses the additive notations, where the symbol 0 denotes the trivial (one element) group.

3.2 On the Homotopy Groups of the Sphere $\mathbb{S}^n$

We start with the simplest case: $\pi_k(\mathbb{S}^n)$ with $0 < k < n$.

Theorem 2. *The kth homotopy group of any higher-dimensional sphere* $\mathbb{S}^n$ *is trivial provided that* $0 < k < n$: $\pi_k(\mathbb{S}^n) = 0$.

Proof. For $k = 1$, the example of the fundamental group of the circle, $\pi_1(\mathbb{S}^1) = \mathbb{Z}$ (proved below, p. 75), shows the necessity of $n > k$. For other values of k the situation is similar when $n = k$, as we will see later.

The triviality of the groups $\pi_{k<n}(\mathbb{S}^n, *)$ is an easy geometrical fact (see Fig. 4.4 in Ch. 4):

According to the Weierstrass Approximation Theorem (see Ch. 4), we can replace the spheroid $\varphi : I^k \to \mathbb{S}^n$ by a homotopic smooth one (defined by polynomials in some coordinates). So we can assume that φ is smooth.

According to Sard's Lemma (p. 20-23), the image set $\varphi(I^k) \subset \mathbb{S}^n$ of a smooth map φ has measure zero, provided that $k < n$. Hence, some point of the sphere is not covered by the image $\varphi(I^k)$ and also some neighbouring n-dimensional disc of this point is free from the points of the spheroid.

Contracting the whole complement of this n-disc to the point $*$, which cannot belong to the disc, we obtain a continuous homotopy of our spheroid to the trivial one $*$ (Fig. 3.4), proving the triviality of the homotopy group:

$$\pi_k(\mathbb{S}^n, *) = 0 \quad (k < n)\,. \qquad \square$$

For $k = n$, the kth homotopy group of $\mathbb{S}^k$ is $\pi_k(\mathbb{S}^k) = \mathbb{Z}$. The integer corresponding to a spheroid homotopy class is just its degree (see Ch. 10).

The homotopy groups of the spheres $\pi_k(\mathbb{S}^n)$, for $k > n$, are (in the general case) still unknown – the computations of these groups is one of the most

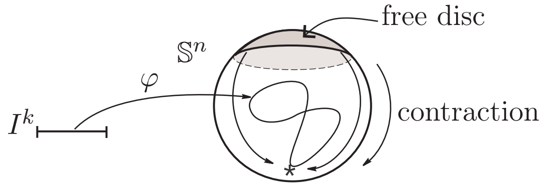


Figure 3.4: Triviality of $\pi_k(\mathbb{S}^n, *)$, for $k < n$.

celebrated unsolved mathematical problems. In 1951, J.-P. Serre succeeded to prove that almost all these groups are finite, with the only exceptions of the groups $\pi_{4\ell-1}(\mathbb{S}^{2\ell})$. For example, $\pi_3(\mathbb{S}^2) = \mathbb{Z}$, $\pi_7(\mathbb{S}^4) = \mathbb{Z} \times \mathbb{Z}_{12}$, $\pi_{11}(\mathbb{S}^6) = \mathbb{Z}$ and $\pi_{15}(\mathbb{S}^8) = \mathbb{Z} \times \mathbb{Z}_{120}$. The answers for the finite groups are known only for rather small values of the difference $k - n$. When this difference reaches hundreds, the necessary computations become too complicated even for the best computers and mathematicians.

It is clear that the homotopy groups of two diffeomorphic manifolds are isomorphic and this isomorphy of homotopy groups remains true for any two homeomorphic topological spaces.

However, in order to have more flexibility when using homotopy groups and to get more powerful theorems, it is convenient to consider an equivalence relation between topological spaces which imposes weaker conditions than homeomorphy, called "homotopy equivalence". Moreover, to compute the homotopy groups, there is a powerful geometric technology called "homotopy lifting". We shall explain and apply these notions in the next section.

3.3 Homotopy and lifting of homotopies

Homotopy and homotopy equivalence. The homotopy between spheroids is a particular case of the following: Given two topological spaces X and Y, a *homotopy* between two maps $\varphi_0 : X \to Y$ and $\varphi_1 : X \to Y$ is a family of homeomorphisms $\{\varphi_s : X \to Y, \quad 0 \leqslant s \leqslant 1\}$ that starts at φ_0 and ends at φ_1. One usually writes $\Phi : X \times [0,1] \to Y, \quad \Phi(x,s) = \varphi_s(x)$.

Example. A curve $\Phi : [0,1] \to Y$ can be considered as a homotopy between its end-points $\varphi_0(\text{point}) = \Phi(0)$ and $\varphi_1(\text{point}) = \Phi(1)$.

In fact, a homotopy between two maps $\varphi_0 : X \to Y$ and $\varphi_1 : X \to Y$ can be considered as a curve that joins them in the space of maps $\{\varphi : X \to Y\}$.

Example. The map $\Phi : \mathbb{R}^n \times [0,1] \to \mathbb{R}^n$ defined by the formula $\Phi(x,s) = \varphi_s(x) = (1-s)x$ is a homotopy between the identity map $\varphi_0(x) = x$ and the map $\varphi_1(x) = O$ in $\mathbb{R}^n$. It "retracts" $\mathbb{R}^n$ to its origin O (Fig. 3.5*a*).

More generally, a *deformation retract* of a space X to its subspace Y is a homotopy $\Phi : X \times [0,1] \to X$ between the identity map $\varphi_0 : X \to X$ and a map $\varphi_1 : X \to X$ such that $\varphi_1(X) \subset Y$ and whose restriction $\varphi_s|_Y : Y \to Y$ is the identity map in Y for all $s \in [0,1]$.

Exercise. Verify that the map $\Phi : (\mathbb{R}^2 \smallsetminus \{O\}) \times [0,1] \to \mathbb{R}^2 \smallsetminus \{O\}$, defined by $\Phi(x,s) = \varphi_s(x) = (1-s(1-1/|x|))\, x$ is a deformation retract of $\mathbb{R}^2 \smallsetminus \{O\}$ to the unit circle (Fig. 3.5*b*).

Homotopy Equivalence. A continuous map $f : M \to \widetilde{M}$ is a *homotopy equivalence* if there exists another continuous map $g : \widetilde{M} \to M$ satisfying:

1) The product $g \circ f : M \to M$ is homotopic to the identity map of M.
2) The product $f \circ g : \widetilde{M} \to \widetilde{M}$ is homotopic to the identity map of $\widetilde{M}$.

If there exist such a homotopy equivalence from X to Y, then X and Y are said to be *homotopy equivalent* (or *homotopic*) and one also says that they have the same *homotopy type*.

Example. The map $\varphi_1 : \mathbb{R}^n \to \mathbb{R}^n$ of the previous example is a homotopy equivalence. Hence $\mathbb{R}^n$ is homotopy equivalent to a point – Fig. 3.5*a*.

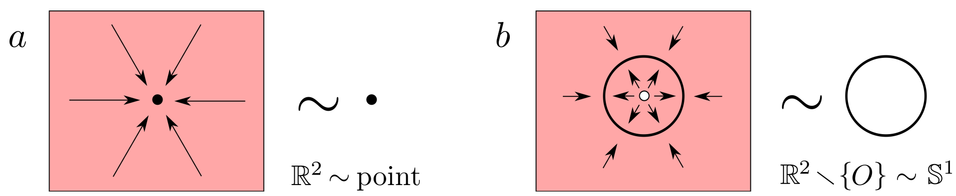


Figure 3.5: The plane is homotopy equivalent to a point; the plane with its origin removed is homotopic to the circle.

Example. The map $\varphi_1 : \mathbb{R}^2 \smallsetminus \{O\} \longrightarrow \mathbb{R}^2 \smallsetminus \{O\}$ of the previous exercise is a homotopy equivalence. Hence $\mathbb{R}^2 \smallsetminus \{O\}$ is homotopic to a circle – Fig. 3.5*b*.

Observe that for two manifolds *homeomorphic* $\Rightarrow$ *homotopic* but the converse is not true: *homotopic* $\not\Rightarrow$ *homeomorphic*. Indeed, the previous examples show that a homotopy can change the dimension. Moreover, a homotopy could not preserve the manifold structure. For example, the two dimensional manifold $\mathbb{R}^2 \smallsetminus \{\text{two different points}\}$ is homotopic to the "eight curve" ∞, which is not a manifold, but just a topological space.

Remark. In fact, the homotopy groups π_k are defined for topological spaces and each homotopy group of a manifold (or topological space) is invariant under homotopy equivalence, that is, *the k-th homotopy groups of any two spaces with the same homotopy type are isomorphic for all $k \in \mathbb{N}$.*

A topological space is said to be *contractible* if it is homotopic to a point. Thus $\mathbb{R}^n$ is contractible and all its homotopy groups are trivial.

Problem. Prove that (or explain why) the following seven manifolds are homotopy equivalent to the "eight curve" ∞: 1. The sphere minus three points; 2. The plane minus two points; 3. The torus minus one point; 4. The Klein Bottle minus one point; 5. The Möbius band minus one point; 6. The 3-space $\mathbb{R}^3$ minus two disjoint lines; 7. The projective plane minus two points.

Homotopy lifting. Let $p : E \to B$ be a continuous map (in most examples that we shall consider it will be a smooth map between two smooth manifolds). Let X be another space or manifold (in most of our examples X will be the cube I^k).

Consider a continuous map $f : X \to E$ and the "projected" continuous map $g = p \circ f : X \to B$, where $g(x) = p(f(x))$, for every $x \in X$ – Fig. 3.6.

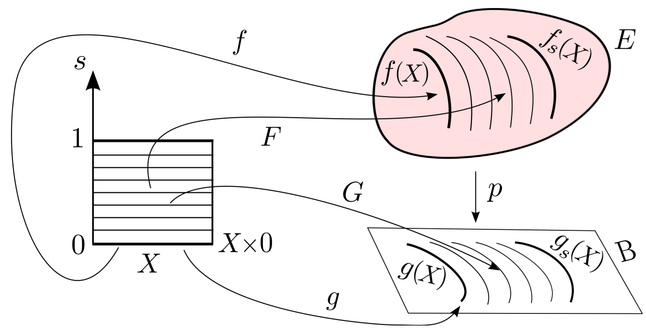


Figure 3.6: A lifting F of the homotopy G of the projection g of the map f.

Suppose there is a homotopy of the projected map g, that is, a continuous family of maps g_s, $0 \leqslant s \leqslant 1$, starting from $g_0 = g$:

$$G : X \times I \to B, \quad G|_{X\times(s\in I)} = (g_s : X \to B).$$

Lifting. A *lifting* of the homotopy G with initial condition f is a homotopy $F : X \times I \to E$ that coincides with f on the fibre $X \times \{0\}$ and is projected to G by p:

$$p(F(x,s)) = G(x,s) \quad \text{for any} \quad x \in X, \ s \in I.$$

Homotopy Lifting Property. A map $p : E \to B$ *has the homotopy lifting property*, if any homotopy G of the projection to B of any initial condition $f : X \to E$ admits a lifting to some homotopy F of the initial condition f.

PROBLEM. Consider the vertical projection map p of the complement $E = \mathbb{R}^2 \setminus 0$ of a point of the plane to the horizontal line $B = \mathbb{R}$ (Fig. 3.7): p sends the point with coordinates (x, y) to the abscissa value, $p(x, y) = x$.

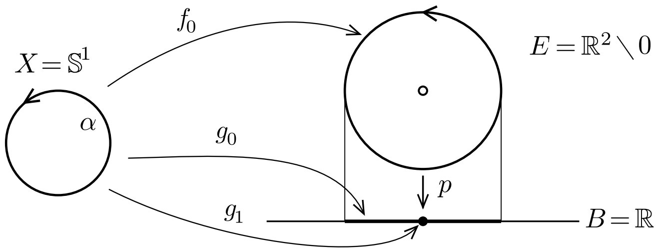


Figure 3.7: An unliftable homotopy $\{g_s\}$ of the projection g_0 of the map f_0.

Has this projection the homotopy lifting property?

SOLUTION. Consider the circle $X = \mathbb{S}^1$ and let the initial condition $f : X \to E$ be the inclusion of the unit circle defined by the formula $f_0(\alpha) = (\cos\alpha, \sin\alpha)$ for the angular coordinate α on $\mathbb{S}^1$.

The projected initial condition is defined by the formula $g_0(\alpha) = \cos\alpha$. The homotopy $G(\alpha, s) = (1 - s)g_0(\alpha)$ cannot be lifted.

Indeed, the polar angular coordinate φ of the point $(x, y) = (r\cos\varphi, r\sin\varphi)$ of E acquires the increment 2π when f_0 makes the turn $0 \leqslant \alpha \leqslant 2\pi$. For any homotopy $F = \{f_s\}$ of f_0 in E the increment of φ should be a continuous function in s, whose value is an integer multiple of 2π because the circle X is closed. Since it is equal to 2π for $s = 0$, it would remain equal to 2π for every moment s of the homotopy $F = \{f_s\}$. Thus, for $s = 1$ it would be also equal to 2π.

However, for $s = 1$, the homotopy $G = \{g_s\}$ downstairs sends the whole circle X to the point $x = 0$ and, hence, the lifted version F ought to send X into the line $x = 0$.

Whence the angular coordinate $\varphi(F(\alpha, 1))$ should be a constant, and its increment along $0 \leqslant \alpha \leqslant 2\pi$ ought to be 0 as well as to be 2π.

This contradiction proves that the homotopy G, for the initial condition f_0, is not liftable.

Problem. Consider the projection of the torus to the circle, sending the point with angular coordinates α, β of $\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$ to the point with angular coordinate α of the first factor-circle, $p(\alpha, \beta) = \alpha$.

Has this projection the homotopy lifting property?

Solution. In the case of the projection of a direct product $E = B \times C$, $p(b, c) = b$, the initial condition is always a pair of maps $f(x) = (g(x), c(x))$ $(x \in X,\ g(x) \in B,\ c(x) \in C)$. Any homotopy G of g is lifted trivially to $F(x, s) = (G(x, s), f(x, 0))$, and hence *the natural projections of direct products always have the homotopy lifting property.* This solves the problem.

The same reasoning proves that the homotopy lifting property holds for the fibration $p : E \to \mathbb{S}^1$ of the Möbius strip into segments and for the fibration of the Klein bottle into circles. Indeed, these fibrations are locally trivial, their difference from the direct product is only global. Restricting the base to a small neighbourhood of a point, we get the direct product structure on the union of the fibres.

Any homotopy can be represented as a product of a finite sequence of local homotopies for which all the points of X remain fixed $(G(x, s) = g(x))$, except for a small neighbourhood of one special point.

This decomposition is explained, for the case $X = I$, in Fig. 3.8.

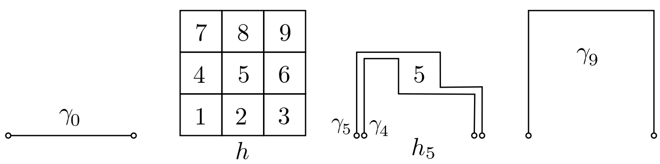


Figure 3.8: The decomposition of the homotopy h into a sequence of 9 local homotopies h_j.

One replaces the homotopy h, sending the big square to M, by the product of 9 homotopies h_j that are non-trivial only on the 9 corresponding squares $(j = 1, \ldots, 9)$. The jth square connects the curves γ_{j-1} and γ_j, and defines a homotopy h_j between the restrictions of h to these two curves.

Quite similarly, for $X = I^2$ one can decompose the 3-cube $X \times I$ of the homotopy G into N^3 small parts and represent the homotopy G as the product of a sequence of N^3 small local homotopies.

A similar reasoning works also for $X = I^k$, for any compact smooth manifold X and in many other cases, including even the applications to the infinite-dimensional objects, like the spaces of loops or of other maps.

We shall use it below for the cases in which the cubes I^k are mapped to the space E of a smooth fibration $p : E \to B$, taking into account sometimes the boundary condition of the spheroids at the point $*$.

Problem. Consider a linear projection of two sides of a triangle to some line (see Fig. 3.9). Under which conditions this projection verifies the homotopy lifting property?

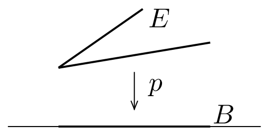


Figure 3.9: Has this projection the homotopy lifting property?

Solution. The homotopy lifting property is satisfied only in the case of the projection of the union of two open sides of the triangle to one of them in the parallel direction to the third side – because it is a trivial direct product.

We will prove that for two closed sides the homotopy lifting property is not verified – see Fig. 3.10.

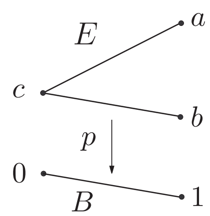


Figure 3.10: A map not satisfying the homotopy lifting property

Let $X = E$ and $f_0 = \mathrm{Id} : X \to E$. Choose the homotopy of the projection $p \circ f_0 : X \to B$ to be the contraction of the segment B into his right extremity. Such a homotopy between the maps $g_0 = p : X \to B$ and $g_1 = X \to \{1\} \subset B$ cannot be lifted.

Indeed, suppose there exists a lifting of the homotopy, $f_t : X \to E$, $t \in [0,1]$. The equalities $g_t(a) = g_t(b) = 1$ imply $f_t(a) = a$ and $f_t(b) = b$. Then the image of the lifted map $f_1 : X \to p^{-1}(1) = \{a\} \cup \{b\}$ must be connected because X is connected. That is, either $f_1(X) = a$ or $f_1(X) = b$.

This contradicts the fact that the extremities of $X \simeq I$ belong to different components of the fibre $p^{-1}(1)$ during all the homotopy process.

In general, the following proposition is deduced from the homotopy lifting property.

Proposition. *If the base B of the projection $E \to B$ is connected, then the different fibres are homotopy equivalent between them.*

In particular, all fibres must have an identical number of connected components. Thus, if we replace the side B of the triangle by the line on which it lies, the homotopy lifting property will not hold even for the projection of the union of two open sides.

The reader can verify that in all other cases some homotopies are not liftable. Some are not liftable even in the open sides case, for which the projection is locally a fibration in some neighbourhood of any point of the manifold B, provided that the direction of the projection is not the direction of the third side of the triangle.

Example. For the projection of the union of two closed sides to one of them along the other, the unliftable homotopy is produced by the motion of the common vertex inside the image side, provided that the initial condition f sends X to the interior part of the second side.

Problem. Consider the linear projections depicted in Fig. 3.11. Which of them verifies the homotopy lifting property?

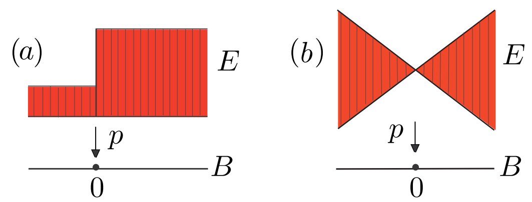


Figure 3.11: Which map satisfies the homotopy lifting property ?

Solution. Projection (a): All fibres are homotopy equivalent and they are all diffeomorphic. However this projection does not verifies the lifting property:

Let X be a point and take the initial condition $f : X \to E$ sending X to the upper part of the fibre over 0. The map $p \circ f : X \to B$ sends X to 0, and any homotopy of this map moving the image to the left cannot be lifted.
Projection (b): The fibres change dimension, but all are homotopy equivalent. The lifting property is verified by this projection.

In order to decide practically whether a given smooth map between two smooth manifolds $p : E \to B$ is locally a trivial fibration, a very useful tool is the implicit function theorem, which claims that locally the level sets form a trivial fibration, provided that the derivative of p is non-degenerate in the sense that the point $e \in E$, where we compute it, is not critical:

$$p_{*e}(T_e E) = T_{p(e)} B \,,$$

for the derivative linear map $p_{*e} : T_e E \to T_{p(e)} B$, defined by the Jacobian matrix of the map p in terms of the local coordinates on E and on B.

In the case where E and B are compact, this implicit function statement suffices to obtain a fibration (or a set of several fibrations in the non-connected case). However, in the non-compact case of Fig. 3.7, there is no critical point, and still the map p is not a fibration: The preimages of different points of the base line in the plane are not homotopy equivalent and the homotopy lifting property fails.

Another example to keep in mind is the projection of a hyperbola in $\mathbb{R}^2$ along its asymptote to a line, say the projection $p(x, y) = x$ of $E = \{(x, y) : xy = 1\}$ to the x-axis B. The projection $p : E \to B$ is a (locally trivial) fibration at every point of the hyperbola, but this projection does not define a fibration of the hyperbola over the straight line, and there exist unliftable homotopies.

This situation is typical in algebraic geometry: The affine manifolds are usually not compact, and the compactness of the projective spaces and of their algebraic submanifolds is quite useful.

Exercise. Let $\overline{M}$ be the compactification in the projective space of the one-sheeted real hyperboloid M, in the affine space $\mathbb{R}^3$, defined by the equation

$$x^2 + y^2 = z^2 + 1.$$

Prove that $\overline{M}$ is a smooth compact submanifold of $\mathbb{R}\mathrm{P}^3$ diffeomorphic to the torus $\mathbb{T} = \mathbb{S}^1 \times \mathbb{S}^1$.

Solution. The factors can be interpreted as the compactifications of the straight "generating" lines a, b on the hyperboloid (Fig. 3.12). We get the product of the two circles $\overline{a}, \overline{b}$ because $\mathbb{R}\mathrm{P}^1 \approx \mathbb{S}^1$.

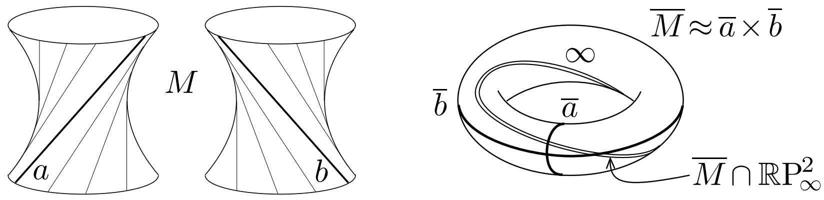


Figure 3.12: The generating lines a and b on the one-sheeted hyperboloid M in $\mathbb{R}^3$, and on its projectivised closure $\overline{M}$ in $\mathbb{R}\mathrm{P}^3 = \mathbb{R}^3 \cup \mathbb{R}\mathrm{P}^2_\infty$.

The intersection of the torus $\overline{M}$ with the "infinitely far" projective plane, added to the affine space $\mathbb{R}^3$ to compactify it, is also a smooth circle (denoted by the symbol ∞ in Fig. 3.12).

3.4 Homotopy exact sequence

Now, suppose that $p : E \to B$ is a fibration or at least that it verifies the homotopy lifting property.

In this case the homotopy groups of the spaces E, B, and of the fibres $F_b = p^{-1}(b)$ (over the points of the base space B) have a remarkable relation. Write $* \in E$ for the chosen basic point of E, choose as the basic point of the base space B the projection of $* \in E$ to B and denote $p(* \in E)$ by $* \in B$. Observe that the fibre $F = p^{-1}(* \in B)$ contains the initial basic point $* \in E$. Consider the homotopy groups π_k of these three spaces at their respective points $* \in E$, $* \in B$ and $* \in F$.

For simplicity we shall suppose these spaces to be connected smooth manifolds, although below we will discuss separately the important case of non-connected fibres occurring, for instance, for the natural double covering $p : \mathbb{S}^n \to \mathbb{R}\mathrm{P}^n$ and for the covering of the circle by the axis of its angular coordinate, $p : \mathbb{R} \to \mathbb{S}^1$.

The inclusion $i : F \to E$ and the projection $p : E \to B$ induce natural homomorphic maps of the homotopy groups,

$$\pi_k(F, *) \xrightarrow{i_*} \pi_k(E, *) \xrightarrow{p_*} \pi_k(B, *).$$

We shall add to these natural short sequences (occurring for every k) a collection of homomorphisms,

$$h : \pi_k(B, *) \to \pi_{k-1}(F, *),$$

connecting these short sequences to form one long (infinite) sequence.

To define the connecting homomorphism h, we use a simple geometric reasoning that generalise the monodromy construction (for more details see Ch. 5, p. 161-164).

Given a k-dimensional spheroid $G : I^k \to B$ (Fig. 3.13), consider it as a 1-parameter family $\{g_s : 0 \leqslant s \leqslant 1\}$ of $(k-1)$-dimensional spheroid-sections, where the cube I^k is considered as the product $I^{k-1} \times I$ and s is the coordinate in the factor I (the parameter space):

$$G(t, s) = g_s(t), \quad t \in I^{k-1}, \ s \in I.$$

This family connects the trivial $(k-1)$-spheroid to itself, $g_0 = g_1 = * \in B$, and each element g_s is an spheroid at the point $* \in B$ because the boundary $\partial(I^{k-1} \times \{s\}) \subset \partial I^k$ is sent to $*$ by G.

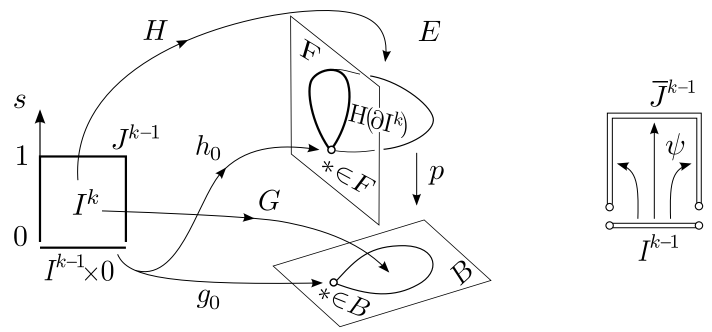


Figure 3.13: Construction of the $(k-1)$-spheroid $H \circ \psi$ in the fibre, from the k-spheroid G in the base of the fibration p.

The trivial spheroid $h_0 = * \in E$ is projected to g_0 by p, and the lifted homotopy H provides a family of spheroids $\{h_s\}$, $0 \leqslant s \leqslant 1$, at the point $* \in E$. In this way, one obtains a map $H : I^{k-1} \times I \to E$ sending the boundary of the cube $I^k = I^{k-1} \times I$ to the fibre $F = p^{-1}(* \in B)$ because H is projected by p to the spheroid G, $p \circ H = G$. In particular, the map H sends the face $I^{k-1} \times \{0\}$ to $* \in F$, but it is not obliged to send all the boundary ∂I^k to the single point $* \in F$, so the remaining part J^{k-1} of ∂I^k may be sent anywhere in the fibre $F = p^{-1}(* \in B)$.

Consider, in the boundary of the cube I^k, the complement to the face $I^{k-1} \times \{0\}$ of that cube, $J^{k-1} = \partial I^k \setminus (I^{k-1} \times \{0\})$. The closure of J^{k-1} is itself homeomorphic to the standard cube I^{k-1}, and can be identified to it by a canonical continuous bijection $\psi : I^{k-1} \to \overline{J}^{k-1}$ – Fig. 3.13.

We have constructed in this way a map of the $(k-1)$-dimensional cube to the fibre F,

$$\widehat{h}(t) = H(\psi(t)), \quad t \in I^{k-1} ,$$

which is essentially the restriction of the lifted homotopy $H = \{h_s\}$ to the complement of the face $I^{k-1} \times \{0\}$ in ∂I^k.

This map $\widehat{h} : I^{k-1} \to F$ is a spheroid in F at the point $* \in F$ because H sends $I^{k-1} \times \{0\}$ to $* \in F$, according to the initial condition of the lifting, and hence $H((\partial I^{k-1}) \times \{0\}) = (* \in F)$.

The coincidence of the boundaries of the face $I^{k-1} \times \{0\}$ and of its complement in ∂I^k, J^{k-1}, implies that $\widehat{h}(t) = (* \in F)$ for any $t \in \partial I^{k-1}$. Thus $\widehat{h}$ is a $(k-1)$-spheroid at $* \in F$.

This is the whole construction of the connecting homomorphism h: We have associated to a k-spheroid G of the base B at the point $* \in B$ a $(k-1)$-spheroid $\widehat{h}$ at the point $* \in F$ of the fibre.

Of course, strictly speaking, a mathematician is obliged to prove two (evident) things more.

First, the homotopy class of the $(k-1)$-spheroid $\widehat{h}$ at $* \in F$ does not depend on the particular choice of the initial spheroid G at $* \in B$ and on the particular choice of the lifting H, it depends only on the homotopy class of G. The formal proof of the intrinsic nature of this construction is an easy new application of the homotopy lifting, applied to the maps of the next cube I^{k+1} rather than of I^k.

Second, one have to check that our construction provides a homomorphism of groups: The $(k-1)$-spheroid $\widehat{h}$ for the product of two k-spheroids G' and G'' is constructed from the product of the $(k-1)$-spheroids $\widehat{h}'$ and $\widehat{h}''$, obtained from G' and from G'', respectively, and is at least homotopic to this product. This statement is a direct corollary of the geometric construction described above, and we leave the details of the proof to the reader.

We have thus constructed an infinite sequence of groups and homomorphisms between the groups:

$$\ldots \to \pi_k(F, *) \to \pi_k(E, *) \to \pi_k(B, *) \to \pi_{k-1}(F, *) \to \ldots \quad . \qquad (1)$$

The main theorem on the homotopy groups of the fibrations is:

Theorem 3. *The sequence* (1) *is exact.*

We recall the exactness definition, from the elementary algebra courses.

Exact and Semi-exact Sequences. A sequence of homomorphisms of groups

$$A \xrightarrow{u} B \xrightarrow{v} C$$

is said to be *exact at B*, if the kernel of the right homomorphism coincides with the image of the left one (Fig. 3.14): $(\operatorname{Im} u = u(A)) = (v^{-1}(0) = \operatorname{Ker} v)$.

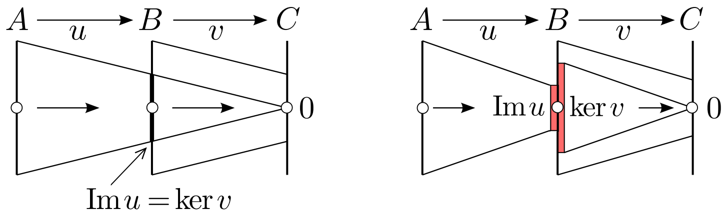


Figure 3.14: Exact and semi-exact sequences.

Using the additive notations, we rewrite the exactness condition in the following form:

$$v \circ u = 0, \quad \text{with} \quad v(b) \neq 0 \quad \text{for} \quad b \quad \text{outside} \quad u(A).$$

The condition $v \circ u = 0$ is called *the semi-exactness condition.* It kills the image of two consecutive homomorphisms of the chain – see the right-hand side picture of Fig. 3.14.

A sequence of homomorphisms having the semi-exactness condition is called *chain complex.* For a chain complex, the image $u(A) \subset \operatorname{Ker} v$ is a subgroup, and the quotient group $\operatorname{Ker} v/\operatorname{Im} u$ is called *the homology group* of the chain (we consider the abelian case here).

The above exactness theorem states the exactness at every place of the sequence.

The proof of this easy theorem is a direct application of the homotopy liftings that makes the explanations extremely difficult; in the same sense that the only way to understand the statements $7 \times 8 = 56$ and $1/2 + 1/3 = 5/6$ is to prove them independently of the teacher, whose proofs might rather make things more difficult. The proverb says that every mathematician should prove these facts by himself once in his life – and perhaps only once because the second time would be equally difficult as the attempts to understand the proof of the teacher. We shall verify just a small part of the exactness.

Proof of the exactness at $\pi_k(E,*)$. Suppose we have an element α of the homotopy group of the fibre $\pi_k(F,*)$.

Including the fibre in the space E of the fibration $p : E \to B$, we get an element $\widehat{\alpha}$ of $\pi_k(E,*)$. Projecting it to the base space, we obtain the trivial element $0 \in \pi_k(B,*)$, since the initial spheroid in F representing α provides a spheroid in E (completely inside F) that represents $\widehat{\alpha}$ and is projected to the point $* \in B$. Thus we have proved the semi-exactness at $\pi_k(E,*)$.

Now, suppose that an element β of $\pi_k(E,*)$ is sent to the trivial element 0 of $\pi_k(B,*)$ by the projection.

The homotopy $G : I^{k+1} \to B$ between the k-spheroid $g_0 : I^k \times \{0\} \to B$, obtained by projecting the initial spheroid $h_0 : I^k \to E$, and the trivial k-spheroid g_1 at $* \in B$ is liftable to a homotopy $H : I^{k+1} \to E$ that includes h_0 in the family of maps $\{h_s\}$ and ends at $s = 1$ with a spheroid h_1 which is projected to $* \in B$ by p. Hence h_1 defines an element α of $\pi_k(F,*)$ and is homotopic to the initial k-spheroid h_0. Thus $\beta = i_*\alpha$ belongs to the image of the left homomorphism of our sequence, $i_* : \pi_k(F,*) \to \pi_k(E,*)$, if it belongs to the kernel of the right one, $p_*\beta = 0$.

Neglecting some (easy) details, this reasoning proves the exactness of the homotopy groups sequence at the group $\pi_k(E,*)$. □

We leave to the reader the pleasure to verify the exactness at the other places of the sequence.

Remark. Strictly speaking, there is something to add in the cases $k = 1$ and $k = 0$, since we have not discussed the group operation in $\pi_0(M,*)$, where one ought to start from the void loops $I^0 \to M$, adding a boundary condition at the empty set ∂I^0.

The best way is to define the 0-dimensional spheroids as the maps φ of the 0-dimensional sphere $\mathbb{S}^0$, which consists of two points noted 0 and 1, to M such that $\varphi(0) = *$. The homotopy classes of such maps describe the pathwise connected components of M. In this way, we define the *set* $\pi_0(M,*)$, but the definition of the *product* of two classes is less clear.

There is a case for which the solution is rather natural: When M is a group and $*$ is the unity element. In this case, the elements of $\pi_0(M,*)$ are the pathwise connected components of the group. Their product is naturally defined by the multiplication of the group elements: The connected component of the product is defined by the connected components of the factors, if the multiplication is continuous.

The exactness holds at the π_1 level too in these cases. We shall see now its practical applications.

3.5 Homotopy Groups of the Circle $\mathbb{S}^1$

Consider the natural fibration $p : \mathbb{R} \to \mathbb{S}^1$, defined by the angular coordinate on the circle. To simplify the fibre description, we divide the angle by 2π. The fibration becomes simply the identification of the circle to the quotient group

$$\mathbb{R}/\mathbb{Z} \approx \mathbb{S}^1 .$$

The fibre $F \subset E$ of our fibration $\mathbb{R} \to \mathbb{S}^1$ is thus the subgroup $\mathbb{Z} \subset \mathbb{R}$ of the integers in the real line E.

The homotopy exact sequence takes the form

$$\ldots \to \pi_k(\mathbb{Z}, 0) \to \pi_k(\mathbb{R}, 0) \to \pi_k(\mathbb{S}^1, *) \to \pi_{k-1}(\mathbb{Z}, 0) \to \ldots .$$

The leftmost and the rightmost groups written above are both trivial for $k > 1$, since 0 is an isolated point of $\mathbb{Z}$ and a spheroid of positive dimension at it is therefore trivial.

Theorem (of algebra)**.** *In an exact sequence of groups*

$$0 \to A \xrightarrow{u} B \to 0$$

the homomorphism u is an isomorphism.

Proof. The exactness at A means that the kernel of u, which coincides with the image of 0, is just the 0 element of A. The exactness at B means that the image of u is the whole group B, since it is sent to 0 by the rightest homomorphism.

Thus u maps A onto B bijectively and is therefore an isomorphism. □

Remark. The same reasoning provides the exactness of the sequence

$$\operatorname{Ker} u \to A \xrightarrow{u} B \to \operatorname{Coker} u ,$$

where the group Coker u is the quotient $B/u(A)$. This sequence can be continued by adding a 0 at the left and a 0 at the right. The exactness of the sequence

$$0 \to K \to A \xrightarrow{u} B \to C \to 0$$

means that $K = \operatorname{Ker} u$ and that $C = \operatorname{Coker} u$.

These remarks of elementary algebra are evident, but useful in the formal calculations. In our case, we deduce from them the conclusion:

$$\pi_k(\mathbb{S}^1) \approx \pi_k(\mathbb{R}), \quad \text{for} \quad k \geqslant 2\,.$$

Of course, for $k = 1$ this conclusion fails, since the fundamental groups are

$$\pi_1(\mathbb{S}^1) = \mathbb{Z}\,, \qquad \pi_1(\mathbb{R}) = 0\,,$$

in accordance with the exactness of the homotopy groups sequence of the fibration $\mathbb{Z} \xrightarrow{i} \mathbb{R} \xrightarrow{p} \mathbb{S}^1$ at $\pi_1(\mathbb{S}^1)$ and with the evident identity $\pi_0(\mathbb{Z}) = \mathbb{Z}$.

Theorem 4. *The homotopy groups of the circle* $\mathbb{S}^1$ *are all trivial, except the fundamental group* $\pi_1(\mathbb{S}^1) = \mathbb{Z}$ *:*

$$\pi_k(\mathbb{S}^1) = 0, \quad \textit{for} \quad k > 1\,.$$

Proof. The homotopy groups of the line $\mathbb{R}$ are trivial: $\pi_k(\mathbb{R}) = 0$. This follows because the interval $|x| < 1$ is both homeomorphic to the line $\mathbb{R}$ and homotopic to its point $x = 0$ by the deformation retract $g_s(x) = (1-s)x$. □

As we have mentioned above, the circle is the only sphere of positive finite dimension for which all the homotopy groups are known.

3.6 Fundamental Groups and Coverings

In the same way as above, the fact that for any $k \geqslant 1$ the k-th homotopy group of a discrete set F_0 is trivial ($\pi_k(F_0, *) = 0$) is used to prove the

Theorem 5. (a) *For all* $k \geqslant 2$ *the homotopy groups* π_k *of the base space* B *of a covering* $E \to B$ *are isomorphic to those of the covering space* E:

$$\pi_k(E, *) \approx \pi_k(B, *) \qquad k \geqslant 2\,;$$

(b) The fundamental group $\pi_1(E)$ of the covering space is isomorphic to a normal subgroup of the fundamental group $\pi_1(B)$ of the base.

Proof. Indeed, since the corresponding exact sequence becomes

$$\begin{aligned} &0 \longrightarrow \pi_k(E,*) \xrightarrow{p_*} \pi_k(B,*) \longrightarrow 0 && \text{for } k \geqslant 2\,,\\ &0 \longrightarrow \pi_1(E,*) \xrightarrow{p_*} \pi_1(B,*) \xrightarrow{\partial} \pi_0(F_0,*) \longrightarrow 0 && \text{for } k = 1\,,\end{aligned}$$

statements (a) and (b) follow directly. □

Problem. Calculate the fundamental group of the eight curve "∞".

Solution. We shall first prove that π_1("∞", $*$) *is non commutative.* Fix an orientation of the circles that form the curve "∞" and denote these oriented loops by a and b – Fig. 3.15. We shall prove that the loops ab and ba are not homotopic. For observe that the liftings of these loops in the 6-fold covering of "∞", $\widehat{a\,b}$ and $\widehat{b\,a}$, have different end-points – Fig. 3.15. If there were a homotopy between ab and ba, it should be liftable to a homotopy between $\widehat{a\,b}$ and $\widehat{b\,a}$ that fixes their end points, but this is impossible.

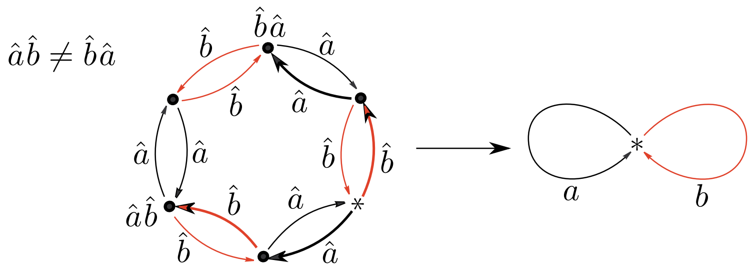


Figure 3.15: The loops ab and ba in the curve "∞" have non homotopic lifts in the covering curve.

Observe that for any non zero integers m, n the powers a^m, b^n are different and non trivial. In fact, π_1("∞", $*$) is the free group F_2 with two generators, namely the loops a and b.

Similarly, the fundamental group of a *bouquet of k circles* (a curve consisting of k circles joined at one common point) is the free group F_k with k generators, each circle being a generator.

These examples show that in general the fundamental group (of a space) is non commutative.

Remark (on free groups). The free group of seven generators F_7 is isomorphic to a normal subgroup of the free group of two generators F_2.

This follows from Theorem 5 and because the 6-fold covering of the curve "∞" is homotopic to a bouquet of 7 circles, whose fundamental group is F_7 (Fig. 3.16). Similarly, using the $(k-1)$-fold covering of "∞" one deducts that *the free group of k generators F_k is isomorphic to a normal subgroup of F_2.*

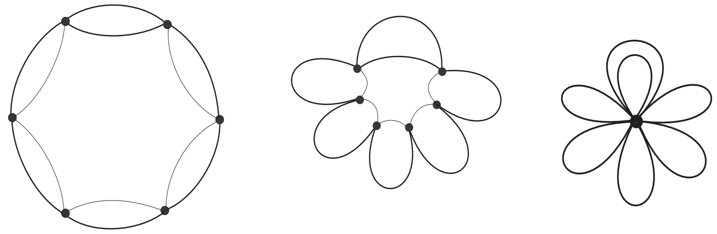


Figure 3.16: The 6-fold covering curve of the "∞" curve is homotopic to the bouquet of 7 circles.

3.7 Homotopy Groups of Real Projective Spaces

We shall extract information from the natural fibration that defines the projective space $\mathbb{R}\mathrm{P}^n$ as the result of gluing the antipodal points of the standard sphere $\mathbb{S}^n$ in Euclidean space $\mathbb{R}^{n+1}$:

$$p : \mathbb{S}^n \to \mathbb{R}\mathrm{P}^n \simeq \mathbb{S}^n/\{\pm 1\}\,.$$

Proposition. *For $k > 1$ the following equality holds $\pi_k(\mathbb{R}\mathrm{P}^n) = \pi_k(\mathbb{S}^n)$.*

Proof. The fibre F over every point of the base $B = \mathbb{R}\mathrm{P}^n$ consists of two opposite points of the sphere. Since the homotopy groups of this fibre are trivial for $k > 1$, the exact sequence of the homotopy groups of our fibration provides the following short sequences, for $k > 1$:

$$0 = \pi_k(F) \to \pi_k(\mathbb{S}^n) \to \pi_k(\mathbb{R}\mathrm{P}^n) \to \pi_{k-1}(F) = 0\,. \tag{1}$$

We conclude that the middle projection defines an isomorphism between the homotopy groups $\pi_k(\mathbb{S}^n)$ and $\pi_k(\mathbb{R}\mathrm{P}^n)$. □

This result, however, does not provide the knowledge of the homotopy groups because the groups $\pi_k(\mathbb{S}^n)$ are in general unknown.

Nevertheless, there is an important partial result.

Theorem 6. *The homotopy groups $\pi_{k>1}$ of the real projective space $\mathbb{R}\mathrm{P}^n$ are trivial, provided that k is smaller than the dimension n of the space:*

$$\pi_k(\mathbb{R}\mathrm{P}^n) = 0, \quad \textit{when} \;\; 1 < k < n\,.$$

Proof. This follows from the just mentioned isomorphism $\pi_k(\mathbb{S}^n) \approx \pi_k(\mathbb{R}\mathrm{P}^n)$, for $k > 1$, and from the triviality of the kth homotopy group of the n-sphere, $\pi_k(\mathbb{S}^n) = 0$, for $k < n$. □

Theorem 7. *For $n > 1$ the fundamental group of $\mathbb{R}\mathrm{P}^n$ is* $\pi_1(\mathbb{R}\mathrm{P}^n) = \mathbb{Z}_2$.

Proof. The exact sequence (1) implies that $\pi_1(\mathbb{R}\mathrm{P}^n) \approx \pi_0(\mathbb{Z}_2) = \mathbb{Z}_2$. □

This means that a real projective line in the real projective plane is not contractible to a point by a continuous homotopy. However, the same line becomes contractible if we consider it as a spheroid making two turns along the circle $\mathbb{R}\mathrm{P}^1 \approx \mathbb{S}^1$.

This particular case can be easily seen geometrically in the following way. Consider the real projective plane as the union of the affine plane $\mathbb{R}^2$ with the line $\mathbb{R}\mathrm{P}^1$ at infinity (which is a circle embedded into $\mathbb{R}\mathrm{P}^2$).

Now consider the neighbourhood of this circle $\mathbb{S}^1$ in $\mathbb{R}\mathrm{P}^2$. It is naturally fibred over $\mathbb{S}^1$, the fibres being small *orthogonal segments* (one can use the metric of the covering sphere $\mathbb{S}^2$ to define the orthogonal segments in $\mathbb{R}\mathrm{P}^2$). The complement to this fibred neighbourhood of the "infinitely far" circle $\mathbb{R}\mathrm{P}^1$ is an ordinary disc in the affine plane $\mathbb{R}^2$ (Fig. 3.17). This fibred neighbourhood is the Möbius band (discovered by Möbius just in this way, as the complement to a neighbourhood of any point of the real projective plane).

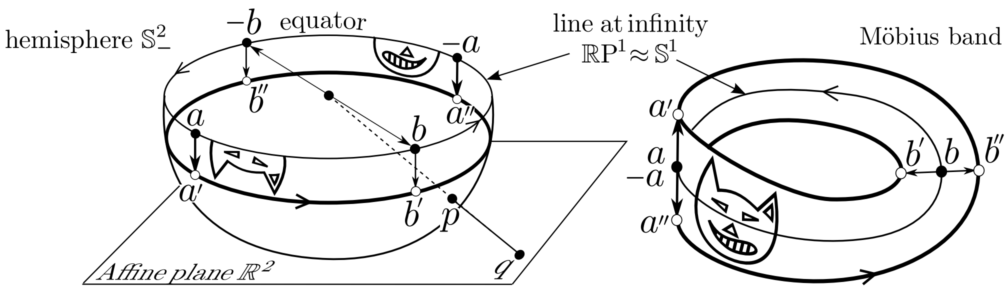


Figure 3.17: The Möbius decomposition of the real projective plane: The complement to a neighbourhood of a point is a Möbius band.

The boundary of the Möbius band is a circle embedded in the real projective plane, it is the boundary of the complementary disc. This circle is embedded in the bigger Möbius bands obtained by taking bigger orthogonal segments, but it is not homotopic there to the central line of the band, which defines a different homotopy class of a map of a circle to the Möbius band.

However, the embedding of the boundary circle is homotopic to the "double embedding" of the central line, that is, an embedding of the circle making two turns around the central line. To obtain that homotopy, one starts at some point of the central line and push along the "positive" normal direction, then one moves along this central line prolonging this pushing continuously. Returning to the initial point after the first turn along the central line, the "positive" normal direction is changed. But continuing the second turn, we return the second time with the initial "positive" normal direction. We have constructed in this way a deformation of the *two-turns map* of $\mathbb{S}^1$ to the central line of the Möbius band.

The resulting deformed closed curve is now contractible in the affine plane $\mathbb{R}^2$ where it bounds a disc. Therefore, for the homotopy class γ of the middle circle ($\mathbb{R}\mathrm{P}^1_\infty$) we find the relation $\gamma^2 = 0$ in $\pi_1(\mathbb{R}\mathrm{P}^2)$.

Similarly, $\pi_1(\mathbb{R}\mathrm{P}^n) = \mathbb{Z}_2$, for any $n \geqslant 2$. This geometric fact is the base of the physical theory of spins of electrons, as we shall explain in a while.

3.8 Homotopy groups of complex projective spaces

We start with the defining fibration

$$p : \mathbb{S}^{2n+1} \to \mathbb{C}\mathrm{P}^n,$$

with circular fibres $\mathbb{S}^1$, using the ordinary definition

$$\mathbb{C}\mathrm{P}^n = \frac{\mathbb{C}^{n+1} \smallsetminus 0}{\mathbb{C} \smallsetminus 0} = \frac{\mathbb{S}^{2n+1}}{\mathbb{S}^1},$$

which associates to a point of $\mathbb{C}\mathrm{P}^n$ the whole circle formed by the complex proportional points of the corresponding point of the unit sphere $\mathbb{S}^{2n+1}$ of $\mathbb{C}^n \approx \mathbb{R}^{2n}$, along which this sphere intersects the complex straight line $\mathbb{C}z$, where $z \in \mathbb{C}^n \smallsetminus 0$.

We get the exact sequence of the homotopy groups:

$$\ldots \to \pi_k(\mathbb{S}^1) \to \pi_k(\mathbb{S}^{2n+1}) \to \pi_k(\mathbb{C}\mathrm{P}^n) \to \pi_{k-1}(\mathbb{S}^1) \to \ldots .$$

For $k > 2$, we already know the triviality of both extremal groups of this fragment of the long exact sequence:

$$\pi_k(\mathbb{S}^1) \approx \pi_{k-1}(\mathbb{S}^1) \approx 0 .$$

Thus we get the nice conclusion that the homotopy groups π_k of the complex projective spaces $\mathbb{C}\mathrm{P}^n$ are isomorphic to the homotopy groups π_k of the spheres $\mathbb{S}^{2n+1}$, if $k > 2$.

Since these last groups are trivial for $k < 2n + 1$, we have proved the

Theorem 8. *The homotopy group $\pi_k(\mathbb{C}\mathrm{P}^n)$ is trivial if $2 < k < 2n+1$.*

For instance, this theorem permits to compute the groups

$$\pi_3(\mathbb{C}\mathrm{P}^2) \approx \pi_4(\mathbb{C}\mathrm{P}^2) \approx 0\,.$$

Moreover, the fundamental group of the complex projective space $\mathbb{C}\mathrm{P}^n$ is trivial, $\pi_1(\mathbb{C}\mathrm{P}^n) = 0$. This is because in the exact sequence it is situated between the trivial groups $\pi_1(\mathbb{S}^{2n+1})$ and $\pi_0(\mathbb{S}^1)$.

The second homotopy group $\pi_2(\mathbb{C}\mathrm{P}^2)$ is not trivial. Indeed, the exact sequence of the homotopy groups

$$\pi_2(\mathbb{S}^5) \to \pi_2(\mathbb{C}\mathrm{P}^2) \to \pi_1(\mathbb{S}^1) \to \pi_1(\mathbb{S}^5)$$

provides the isomorphism

$$\pi_2(\mathbb{C}\mathrm{P}^2) \approx \mathbb{Z}\,,$$

since both groups $\pi_2(\mathbb{S}^5)$ and $\pi_1(\mathbb{S}^5)$ are trivial.

The generating map of the sphere $\mathbb{S}^2$ to the complex projective plane is provided by the inclusion of the complex projective line $\mathbb{C}\mathrm{P}^1 \approx \mathbb{S}^2$ to $\mathbb{C}\mathrm{P}^2$.

In the real case such a spheroid will become homotopic to the trivial one if we cover it twice. However, our calculations show that in the complex case the situation is different: The sphere inclusion γ (for instance, on the complex projective line at infinity of the complex projective plane $\mathbb{C}\mathrm{P}^2$, which one has to add to the affine plane $\mathbb{C}^2$ to obtain the projective one) remains uncontractible even in its multiple versions: $k\gamma \neq 0$ in $\pi_2(\mathbb{C}\mathrm{P}^2, *)$.

3.9 Homotopy groups of the groups of rotations

To study just one connected component of the *orthogonal group* $\mathrm{O}(n)$ of isometries of Euclidean vector space $\mathbb{R}^n$, we fix an orientation of $\mathbb{R}^n$ and consider the group of rotations $\mathrm{SO}(n)$, called *special orthogonal group*. It is also the group of orthogonal $n \times n$ matrices of determinant 1.

The group of rotations of Euclidean plane, $\mathrm{SO}(2)$, is simply the circle $\mathbb{S}^1$, for which we have already calculated its homotopy groups (p. 75).

The group of rotations of Euclidean 3-space, $\mathrm{SO}(3)$, is already very interesting, and we shall now study the topology of this 3-dimensional manifold.

Theorem 9. *The group of rotations* $\mathrm{SO}(3)$ *is diffeomorphic to the space of the tangent vectors of length* 1 *of the sphere* $\mathbb{S}^2$.

Proof. Fix in Euclidean space an orthonormal frame (that is, of three orthogonal vectors u, v, w of length 1). Any rotation $g \in \mathrm{SO}(3)$ sends it to a similar frame (gu, gv, gw), defining the same orientation of Euclidean space $\mathbb{R}^3$. We can therefore identify the manifold $\mathrm{SO}(3)$ with the manifold M^3 of all orthonormal frames (u, v, w) that define the standard orientation of Euclidean space $\mathbb{R}^3$.

The first vector of the frame belongs to the unit sphere $\mathbb{S}^2$. The second vector, being orthogonal to the first one, becomes a tangent vector of the sphere at the end point u of the first vector, if we translate it from the origin to this point of the sphere by parallel translation (Fig. 3.18).

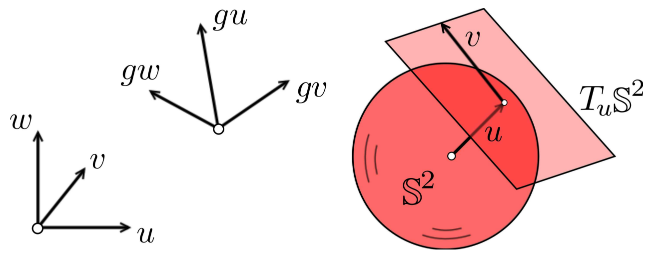


Figure 3.18: Identification of the rotations of the 3-space to the unit tangent vectors of a sphere.

The third vector is well defined by the first two, since it is orthogonal to both of them, and its sign is dictated by the orientation condition.

Thus we have constructed the diffeomorphisms

$$\mathrm{SO}(3) \to M^3 \to T_1\mathbb{S}^2\,,$$

proving the theorem. $\square$

3.9.1 Stabilistion of the Homotopy Groups $\pi_k(\mathrm{SO}(n))$

The above theorem can be also interpreted in terms of the natural "first vector" fibration $p : \mathrm{SO}(3) \to \mathbb{S}^2$, with fibre $\mathbb{S}^1$ (in the preceding notations the projection is defined by the formula $p(u, v, w) = u$).

In the same way, for the manifold of rotations of Euclidean $(n+1)$-space $\mathrm{SO}(n+1)$ we construct the analogous "first vector" fibration:

$$p : \mathrm{SO}(n+1) \to \mathbb{S}^n\,.$$

Its fibre is the space of the (correctly oriented) frames in the tangent space of the n-dimensional sphere, $T_u\mathbb{S}^n$, being therefore diffeomorphic to the group $\mathrm{SO}(n)$. Hence, the exact sequence of the homotopy groups associated to this fibration consists of the following fragments

$$\ldots \pi_{k+1}(\mathbb{S}^n) \to \pi_k(\mathrm{SO}(n)) \to \pi_k(\mathrm{SO}(n+1)) \to \pi_k(\mathbb{S}^n) \to \ldots\ .$$

For $k+1 < n$, the groups in both extremes of these fragments are trivial. Therefore, we get the isomorphism relation $\pi_k(\mathrm{SO}(n)) \approx \pi_k(\mathrm{SO}(n+1))$ and we get the "stable range" for $\pi_k(\mathrm{SO}(n))$:

Theorem 10. *The kth homotopy group $\pi_k(\mathrm{SO}(n))$ stabilises for large values of n, namely, it becomes independent of n whenever $k \leqslant n-2$:*

$$\pi_k(\mathrm{SO}(k+2)) \approx \pi_k(\mathrm{SO}(k+3)) \approx \ldots \approx \pi_k(\mathrm{SO}(\infty))\,.$$

The last symbol is simply an abbreviated notation for this "stabilised homotopy group".

Example. The fundamental groups ($k=1$) stabilise of at $n=3$:

$$\pi_1(\mathrm{SO}(3)) \approx \pi_1(\mathrm{SO}(4)) \approx \ldots\ .$$

3.9.2 Fundamental Group of the manifold $\mathrm{SO}(n)$

In view of the previous example, it becomes interesting to find the fundamental group of the group of rotations $\mathrm{SO}(3)$ of Euclidean 3-space. The following description of the rotations provides a geometric way to find it.

Note that every rotation of the 3-space has a rotation axis spanned by an eigenvector with eigenvalue 1. Choose it to have length one and denote it v.

To prescribe the rotation, knowing the rotation axis, one should fix the angle ϑ of rotation around this axis. Multiplying v by this angle ϑ (and taking into account that, whatever be v, we approach the identity map as $\vartheta \to 0$), we get a topological description of the manifold M formed by the elements of the group $\mathrm{SO}(3)$ as the 3-space of the vectors $\vartheta\, v$.

This description is, however, not a bijective homeomorphism of M to $\mathbb{R}^3$, since the angle of rotation ϑ is defined by the rotation only up to the addition of 2π. We may therefore restrict the angles to the domain $|\vartheta| \leqslant \pi$, modelling M by the ball of radius π in Euclidean space $\mathbb{R}^3$.

This model is still not bijective because the rotations around two opposite directions v and $-v$ are the same. To avoid this difficulty, we restrict the angle of rotation to the interval $0 \leqslant \vartheta \leqslant \pi$. But even in this case there remains the ambiguity for $\vartheta = \pi$ because the rotation of the angle π around the vector v coincides with the rotation of the angle π around $-v$.

Thus we get a homeomorphic model of SO(3) from the closed ball of radius π in Euclidean 3-space, by gluing together the opposite points of the sphere bounding this ball. This homeomorphism provides an important description of the topology of the special orthogonal group.

Theorem 11. *The group of rotations of Euclidean* 3*-space,* SO(3)*, is homeomorphic to the real projective space of dimension* 3*:* $\mathrm{SO}(3) \approx \mathbb{RP}^3$.

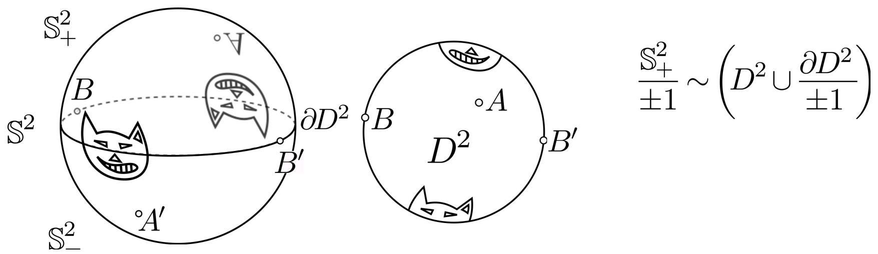


Figure 3.19: Topological model of the projective plane as a disc with glued opposite points of its boundary, $B \sim B'$.

Proof. The construction of the projective space $\mathbb{RP}^3$ from the 3-sphere, by gluing the opposite points, may be replaced by the restriction to the northern hemisphere of $\mathbb{S}^3$ remaining only the gluing of the opposite points on its equatorial boundary $\mathbb{S}^2$ (see Fig. 3.19 for the similar construction of the projective plane $\mathbb{RP}^2$ from the northern hemisphere of $\mathbb{S}^2$ by gluing the opposite points of its equator $\mathbb{S}^1$).

Thus the construction identifies $\mathbb{RP}^3$ with the closed 3-ball whose opposite points on its boundary are glued, and the preceding description of the rotations in terms of v and ϑ provides the proof of Theorem 11. $\square$

The fact that these two smooth manifolds are not only homeomorphic but also diffeomorphic follows from the theory of spins described below.

Corollary. *The group of rotations of Euclidean* 3*-space satisfies*

$$\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2 \quad \textit{and} \quad \pi_2(\mathrm{SO}(3)) = 0\,.$$

Proof. By Theorem 11, $\pi_k(\mathrm{SO}(3)) = \pi_k(\mathbb{R}\mathrm{P}^3)$. We have already proved (p. 77) that $\pi_1(\mathbb{R}\mathrm{P}^n) = \mathbb{Z}_2$ for $n > 1$ and $\pi_k(\mathbb{R}\mathrm{P}^n) = 0$ if $0 < k < n$. □

Corollary. *The fundamental group of the group of rotations of the Euclidean space of dimension greater than* 2 *is the* 2*-elements group* $\mathbb{Z}_2$:

$$\pi_1(\mathrm{SO}(n)) \approx \mathbb{Z}_2, \quad \textit{for} \;\; n \geqslant 3.$$

Proof. Indeed, all these fundamental groups are isomorphic to the fundamental group of SO(3), which is $\mathbb{Z}_2$ by the preceding corollary. □

Remark. The physical way to describe this situation is to introduce the so-called *spin groups*, which we shall now briefly describe.

3.10 Universal covering spaces and spin groups

Let M be a manifold whose fundamental group is non-trivial: $\pi_1(M) \neq 0$. One can cover M by a simply connected* manifold $\widehat{M}$, in a similar way as the coordinate line $\mathbb{R}$ of the angle variable covers the circle $\mathbb{S}^1$, and as the plane $\mathbb{R}^2$ (of the two angular coordinates on the torus) covers the torus $\mathbb{T}^2$.

This useful construction starts from the space of the continuous maps $\varphi : [0,1] \to M$ that send the initial point to $*$ (i.e. $\varphi(0) = *$), while the image of the end-point, $x = \varphi(1)$, is unrestricted. Then one constructs the manifold consisting of all homotopy classes of such paths with fixed end-point $\varphi(1) = x$, that is, each class includes all homotopic paths φ that join $\varphi(0) = *$ with a fixed end-point $x = \varphi(1) \in M$.

Denote by $\widehat{M}$ the resulting set of equivalence classes. It is naturally projected to M by associating to the path φ its end-point $x = \varphi(1)$. This "projection" map is locally a homeomorphism because a path sufficiently close to φ is defined by its end-point up to a homotopy in the class of the paths with the same end-point, and hence a neighbourhood of the equivalence class of φ in $\widehat{M}$ is naturally identified to a neighbourhood of x. Thus $\widehat{M}$ is a manifold if M is a manifold, and the difference between $\widehat{M}$ and M can only be global, like the difference between the circle $M = \mathbb{S}^1$ and the line $\widehat{M} = \mathbb{R}$ that covers the angle coordinate.

Universal covering space. This manifold $\widehat{M}$ as well as any manifold diffeomorphic to it is called *the universal covering space of* M.

*A manifold is called *simply connected* if its fundamental group is trivial.

Proposition. *The fundamental group of the universal covering $\widehat{M}$ is trivial.*

Proof. To prove that any loop $\widehat{\gamma}$ in $\widehat{M}$ is homotopic to the trivial loop $\widehat{*}$ in $\widehat{M}$, it suffices to prove that its projection γ to M is homotopic to the trivial loop $*$ in M (by the Lifting Homotopy Property for coverings).

We represent the points $\widehat{\gamma}(t)$ of $\widehat{\gamma}$ by concrete paths $\varphi_t : [0,1] \to M$ that vary continuously as we move along $\widehat{\gamma}$ (of course $\varphi_t(0) = *,\ \varphi_t(1) = \gamma(t)$).

This continuous 1-parameter family of paths defines a continuous map $\Gamma : I \times I \to M$ with $\Gamma(t,s) = \varphi_t(s)$ – Fig. 3.20-left. Since $\varphi_0(s) = \varphi_1(s) = *$, our map Γ is also a continuous 1-parameter family $\{g_s : s \in [0,1]\}$ of loops (Fig. 3.20-right):

$$\Gamma(t,s) = g_s(t), \quad t \in I,\ s \in I.$$

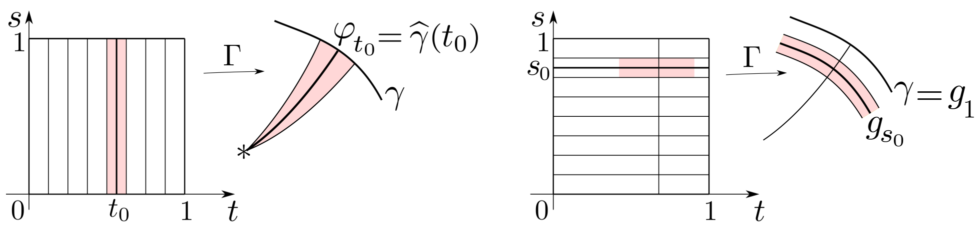


Figure 3.20: Left: the continuous map $\Gamma : I \times I \to M$ determined by the family of paths $\{\varphi_t\}$ given by $\widehat{\gamma}$. Right: the family of loops $\{g_s\}$ determined by Γ.

This family connects the loop $g_1 = \gamma$ to the constant loop $g_0 = * \in M$. Hence γ is is homotopic to the trivial loop $*$. In consequence $\pi_1(\widehat{M}) = 0$. $\square$

In other terms, each path φ_t is continuously deformed to the trivial path along itself, defining shorter paths during the deformation. This generates both a homotopy of $\widehat{\gamma}$ to a new loop in $\widehat{M}$ consisting of the shortened paths and a new map Γ_λ – Fig. 3.21. At the end of the homotopy the deformed loop will become the trivial loop $\widehat{*}$ in $\widehat{M}$.

Remark. Equivalently, we can define a *universal covering* of M as a simply connected manifold covering M, and then prove that all such covering manifolds are diffeomorphic.

Example. The 2-fold covering defining the projective n-space

$$\mathbb{S}^n \to \mathbb{RP}^n \quad (n > 1)$$

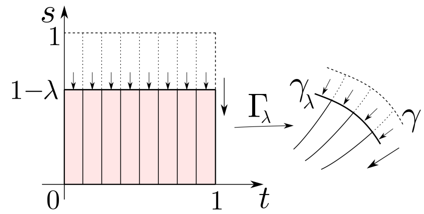


Figure 3.21: The homotopy of $\widehat{\gamma}$ obtained by deforming each path φ_t along itself.

is isomorphic to the universal covering of $M = \mathbb{R}\mathrm{P}^n$ that we have constructed above for a general M: $\mathbb{S}^n \approx \widehat{\mathbb{R}\mathrm{P}^n}$. Note that the universal covering space $\mathbb{S}^n$ is not contractible in this example.

Let M be a non simply connected manifold. Suppose now that M is a group with continuous operation, and that the basic point $*$ is the neutral element ("unity") of this group.

In this case we can define a group operation also in the universal covering space $\widehat{M}$. It is unambiguously defined by the multiplication of the elements of the paths φ and ψ that represent the two points of the covering space $\widehat{M}$: The products $\varphi(t)\psi(t)$ define a continuous path starting at $*$, and whose homotopy class (in the space of the paths starting at $*$ and ending at a fixed point of M) does not depend on the choice of the representatives φ and ψ, but only on their homotopy classes.

In this way, we have defined the multiplication $[\varphi][\psi] = [\varphi\psi]$ of the homotopy classes and hence of the points of the covering space $\widehat{M}$. This operation makes from the space $\widehat{M}$ a group that is locally isomorphic to the initial group M, and then the natural projection $p : \widehat{M} \to M$ is a group homomorphism.

So we obtain from the non simply connected groups $\mathrm{SO}(n)$ their simply connected covering groups. In the cases $n > 2$ these coverings are 2-fold coverings and the covering groups are called *spin groups*,

$$\mathrm{Spin}(n) = \widehat{\mathrm{SO}(n)}\,.$$

We get homomorphic two-fold coverings with kernel $\mathbb{Z}_2$ (of two elements)

$$\mathbb{Z}_2 \to \mathrm{Spin}(n) \to \mathrm{SO}(n) \qquad (n > 2)\,.$$

For the case $n = 3$ the covering manifold is the 3-sphere, and we get a group fibration with fibre $\mathbb{Z}_2$,

$$\mathbb{Z}_2 \to \mathbb{S}^3 \to \mathrm{SO}(3)\,,$$

which associates to each rotation of Euclidean 3-space two opposite points of the 3-sphere, which is equipped with a remarkable group structure.

This group was discovered by Hamilton, who had tried for many years to find a description of the rotations of the 3-space, similar to the description of the plane rotations by the complex numbers – due to Wessel, a Dutch mathematician from whom Abel learnt it.

Finally, Hamilton invented the necessary generalisation of the theory of complex numbers – his quaternion theory that we shall now shortly explain.

3.11 The group of quaternions

The starting point is the use of the complex numbers for the description of the rotations of the plane.

The argument of the product of two complex numbers is equal to the sum of their arguments. So the multiplication of all complex numbers w by a fixed complex number z whose modulus equals 1 is the rotation of the w-plane by the angle $\alpha = \arg z$, that is, $z = \cos\alpha + i\sin\alpha$.

The simple multiplication rule of the complex numbers replaces a lot of long formulae of trigonometry.

For instance, denoting by β the argument of w and supposing its modulus to be also 1, $w = \cos\beta + i\sin\beta$, we get $zw = \cos(\alpha+\beta) + i\sin(\alpha+\beta)$.

So the obvious multiplication of complex numbers

$$\begin{aligned} zw &= (\cos\alpha + i\sin\alpha)(\cos\beta + i\sin\beta) \\ &= (\cos\alpha\cos\beta - \sin\alpha\sin\beta) + i(\cos\alpha\sin\beta + \sin\alpha\cos\beta)\,. \end{aligned}$$

provides the non-obvious trigonometric identities:

$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta, \quad \sin(\alpha+\beta) = \cos\alpha\sin\beta + \sin\alpha\cos\beta\,.$$

The problem is now to extend these formulae to the rotations in the 3-dimensional space.

Hamilton's idea had been to find a similar simple algebra for the description of the rotations of the 3-space. After a long search, the 16th October 1843, while walking with his wife along the Royal Canal to a meeting of the Royal Irish Academy in Dublin, he discovered the desired formula.

The science historians describe his success by the alcohol influence in that mouldy weather evening of autumn. Anyway, he carved with a knife his new-discovered formula on a stone of the Brougham Bridge over the Royal Canal that he was then crossing. Regrettably, no trace of the engraved formula stays and a plaque erected by the Royal Irish Academy in 1958 marks the spot today (sometimes covered with graffiti).

Hamilton even wrote a poem about quaternions:

The Tetractys

Or high Mathesis, with her charm severe,
Of line and number, was our theme; and we
Sought to behold her unborn progeny,
And thrones reserved in Truth's celestial sphere:
While views, before attained, became more clear;
And how the One of Time, of Space the Three,
Might, in the Chain of Symbol, girdled be:
And when my eager and reverted ear
Caught some faint echoes of an ancient strain,
Some shadowy outlines of old thoughts sublime,
Gently he smiled to see, revived again,
In later age, and occidental clime,
A dimly traced Pythagorean lore,
A westward floating, mystic dream of four.

We shall not follow Hamilton's way to guess this result. As Sylvester explained, the proofs of the general statements are easier than those of their particular cases, in spite of the fact that these particular cases are included in the general statements. Sylvester's remark is the philosophical base of the Bourbakisation of the modern presentations of mathematics. So, we start with the axioms of Hamilton.

To describe the rotations of the 3-space, he used a 4-dimensional generalisation of the complex numbers that he called *quaternions*.

The quaternions form a 4-dimensional real vector space $\mathbb{H} = \mathbb{R}^4$. Its four basic vectors are traditionally denoted by the symbols $1, i, j, k$. So any quaternion is a linear combination with real coefficients of the 4 basic vectors:

$$q = a1 + bi + cj + dk\,. \tag{1}$$

To construct the Hamilton algebra of the quaternions, we have to define its two operations. The addition of quaternions is provided by the vector space structure of $\mathbb{R}^4$: One simply adds the coordinates of the two addends.

The multiplication operation is bilinear and hence we need only to define the multiplication table of the basic vectors, similar to the table for the complex numbers: $1 \cdot 1 = 1$, $1 \cdot i = i \cdot 1 = i$, $i \cdot i = -1$.

The formula curved by Hamilton at Brougham bridge is equivalent[*] to

$$ij = k\,,$$

which is the main part of the table containing also the following relations:

$$1 \cdot i = i \cdot 1 = i, \quad 1 \cdot j = j \cdot 1 = j, \quad 1 \cdot k = k \cdot 1 = k, \quad i^2 = j^2 = k^2 = -1\,.$$

For the multiplication of the three elements i, j, k one uses their alphabetic cyclic order. The product of two subsequent elements is always, by definition, the third one:

$$ij = k, \quad jk = i, \quad ki = j\,.$$

An important difference with the complex numbers is the antisymmetry of the multiplication of these units:

$$ji = -k, \quad kj = -i, \quad ik = -j\,.$$

These defining rules end the multiplication table of Hamilton.

To understand better the multiplication, let us decompose the quaternion (1) into its *scalar part* $a \cdot 1$, usually denoted simply by a, and its remaining *vector part* (or *purely imaginary part*)

$$\vec{v} = bi + cj + dk \in \mathbb{R}^3\,.$$

Multiplying two quaternions

$$q = a + \vec{v}\,, \quad q' = a' + \vec{v}\,'\,,$$

we obtain from the multiplication table the following expressions for the scalar and vector parts of the product:

$$qq' = (aa' - \langle \vec{v}, \vec{v}\,' \rangle) + (a\vec{v}\,' + a'\vec{v} + [\vec{v}, \vec{v}\,'])\,. \tag{2}$$

The symbol $\langle \vec{v}, \vec{v}\,' \rangle$ means here the sum of the products of the coordinates $bb' + cc' + dd'$, and the minus sign of this term is explained by the axiom $i^2 = j^2 = k^2 = -1$. One usually calls $\langle \vec{v}, \vec{v}\,' \rangle$ the *scalar product* of the vectors $\vec{v}$ and $\vec{v}\,'$.

[*] He used rather the form $ijk = -1$.

The symbol $[\vec{v}, \vec{v}\,']$ stands for the vector whose coordinates are provided by the above multiplication table. For instance, its k-component is equal to $bc' - b'c$ (since $ij = -ji = k$). It is usually called *vector product* of the vectors $\vec{v}$ and $\vec{v}\,'$ in the oriented Euclidean 3-space (generated by i, j, k).

In fact the origin of the terms "scalar product" and "vector product" is Hamilton's formula (2).

Notice that the multiplication of two quaternions is commutative if one of them is real.

The main theorem, which explains the use of the quaternions for the description of the rotations of the 3-space, is the following result of Hamilton.

In Euclidean space $\mathbb{R}^3$ with orthonormal basis i, j, k, associate to the rotation of angle ϑ around the vector $\vec{w}$ of length 1 the quaternion

$$q = \cos(\vartheta/2) + \sin(\vartheta/2)\vec{w}\,. \tag{3}$$

Hamilton Homomorphism Theorem. *The product of rotations is represented by the Hamilton product of the quaternions: The rotation associated to the quaternion qq' by formula* (3) *is the product of the rotation g associated to the quaternion q, with the rotation g' associated to the quaternion q':*

$$(q \to g, \quad q' \to g') \implies (qq' \to gg')\,.$$

The proof of this important homomorphism theorem and that of formula (3) are given below.

The remarkable difference between formula (3) for the quaternionic description of the 3-space rotations and the corresponding formula for the description of the plane rotations by the complex numbers is the presence of the denominator 2 under the rotation angle ϑ. This number 2 has a topological reason.

The angle of rotation ϑ is determined by the rotation g only up to addition of an integral multiple of 2π. By this ambiguity the angle $\vartheta/2$ is defined only up to addition of an integral multiple of π, whose addition may replace the quaternion q by $-q$.

Thus the relation (3) between quaternions and rotations is not bijective: Two opposite quaternions q and $-q$ describe the same rotation $g \in \mathrm{SO}(3)$.

The denominator 2 in formula (3) is called "Rodrigues formula". It seems that Rodrigues, as well as some other mathematicians, was aware of the Hamilton type description of the rotations before Hamilton published it.

Anyway, the quaternionic description is very useful because it is much easier to multiply quaternions than to work with the three Euler angles or with the nine elements of the 3×3 orthogonal matrices. At present all controls of satellite rotations (both in Russia and in USA) are performed by quaternionic algebra computations.

To prove the Hamilton Homomorphism Theorem, we start with some generalities.

Quaternion Norm. The *norm* $|q|$ of a quaternion $q = a + bi + cj + dk$ is the square root of the sum of the squares of the 4 coefficients:

$$|q|^2 = a^2 + b^2 + c^2 + d^2 \,.$$

In other words, it is the Euclidean length of the vector q in Euclidean space $\mathbb{R}^4$ (of the quaternions).

The quaternion *conjugated to* $q = a + bi + cj + dk$ is

$$\overline{q} = a - bi - cj - dk \,,$$

that is, $\overline{q} = a - \vec{v}$ for $q = a + \vec{v}$.

Obviously, $|q|^2 = q\overline{q}$, and hence

$$q^{-1} = \frac{\overline{q}}{|q|^2} \,, \tag{4}$$

similarly to the theory of complex numbers.

Proposition. *The norm of the product of two quaternions equals the product of their norms.*

Proof. First observe that for the conjugation of the product of two quaternions we have $\overline{QR} = \overline{R}\,\overline{Q}$. This equality follows from formula (4) because $(QR)^{-1} = R^{-1}Q^{-1}$: One puts on the coat after the shirt, but one has to put it down at first. Thus we get

$$|qq'|^2 = (qq')(\overline{qq'}) = qq'\overline{q'}\overline{q} = q|q'|^2\overline{q} = |q'|^2 q\overline{q} = |q'|^2|q|^2 \,. \qquad \square$$

Problem (In number theory). Let m be an integer that is the sum of four squares of integers and m' an integer that is also representable as the sum of four squares of integers. Prove that their product mm' is also representable in this way.

Hint. Multiply the quaternions q and q' whose squared norms are m and m'.

Corollary. *The multiplication of all the quaternions by a quaternion of norm* 1 *is an orthogonal transformation of the Euclidean space of quaternions* $\mathbb{R}^4$ (be it the multiplication from the left or from the right):

$$(q\cdot) \in \mathrm{SO}(4) \quad \textit{and} \quad (\cdot q) \in \mathrm{SO}(4) \quad \textit{if} \quad |q| = 1\,.$$

Consider now the following operation that is associated to any quaternion q of norm 1 and acts on the space of quaternions $\mathbb{H} = \mathbb{R}^4$:

$$A_q : \mathbb{H} \to \mathbb{H}, \quad A_q(h) = qhq^{-1}\,.$$

Theorem 12. *The linear operator $A_q : \mathbb{R}^4 \to \mathbb{R}^4$ acts as a rotation on Euclidean* 3*-space* $\mathbb{R}^3$ *of the purely imaginary (vectorial) quaternions.*

Proof. The vector 1 of $\mathbb{R}^4$ is preserved, since $A_q 1 = qq^{-1} = 1$. The purely imaginary (vectorial) orthogonal complement $\mathbb{R}^3$ to 1 is also invariant, since the transformation A_q is orthogonal. The restriction g of A_q to $\mathbb{R}^3$ preserves the distances, since $\mathbb{R}^3$ is a part of $\mathbb{R}^4$.

The orientation of this subspace $\mathbb{R}^3$ is also preserved, since the quaternions q of norm 1 form the sphere $\mathbb{S}^3$ in $\mathbb{R}^4$, which is connected.

The transformation $g : \mathbb{R}^3 \to \mathbb{R}^3$ belongs therefore to the group of rotations SO(3). □

We shall now prove the basic formula (3).

Decomposing the quaternion q of norm 1 into its real and purely imaginary parts, $q = a + \vec{v}$, we observe that $a^2 + \langle \vec{v}, \vec{v} \rangle = 1$, and hence we can represent the addends as $a^2 = \cos^2\alpha$, $|v|^2 = \sin^2\alpha$ for some angle α. Then we write the vector $\vec{v}$, of squared length $\sin^2\alpha$, as the product $\vec{v} = (\sin\alpha)\vec{w}$, with a normalised vector $\vec{w}$ proportional to $\vec{v}$, but having length 1. In this way we obtain the expression

$$q = c + s\vec{w} \qquad (c = \cos\alpha, s = \sin\alpha)\,.$$

Such quaternions q with a fixed $\vec{w}$ and variable angle α form a circle $\mathbb{S}^1$ in the space of quaternions, which is a one-parameter subgroup parametrised by α of the group of quaternions of norm 1. The formula $q(\alpha)q(\beta) = q(\alpha+\beta)$ is just the formula of the addition of arguments of the complex numbers.

Proposition. *The rotation* $g = A_q|_{\mathbb{R}^3}$ *associated to a quaternion* $q = c + s\vec{w}$ *of norm* 1 *preserves the vector* $\vec{w} \in \mathbb{R}^3$.

Proof. We have $\overline{q} = c - s\vec{w}$, and hence, by the multiplication formula (2),

$$\begin{aligned}
q\vec{w}q^{-1} &= (c + s\vec{w})\vec{w}(c - s\vec{w}) \\
&= (c\vec{w} - s\langle\vec{w}, \vec{w}\rangle + s[\vec{w}, \vec{w}])(c - s\vec{w}) \\
&= (-s + c\vec{w})(c - s\vec{w}) \\
&= (-sc + cs)\langle\vec{w}, \vec{w}\rangle + c^2\vec{w} + s^2\vec{w} - cs[\vec{w}, \vec{w}] \\
&= 0 + 1 \cdot \vec{w} - 0 \\
&= \vec{w}\,.
\end{aligned}$$

□

Proposition. *The angle* ϑ *of the rotation* g *around the axis* $\vec{w}$ *equals* 2α.

This may be seen directly from formula (2), but instead of this trigonometric calculation, we prefer a more geometric way to explain it. We shall first prove that the map sending the quaternions $q \in \mathbb{S}^3$ of length 1 to the group of rotations SO(3) is a group homomorphism. This is the Hamilton Homomorphism Theorem formulated above.

Indeed, according to the definition of the operators A_q, we have

$$\begin{aligned}
A_{qq'}h &= qq'h(qq')^{-1} = qq'h(q')^{-1}q^{-1} = q(q'h(q')^{-1})q^{-1} = q(A_{q'}h)q^{-1} \\
&= A_q(A_{q'}h)\,.
\end{aligned}$$

Thus the Hamilton Homomorphism Theorem is proved.

So the set of quaternions q associated to a fixed vector $\vec{w}$ with arbitrary values of the angle α produces the one-parameter commutative group of the rotations g around the axis $\vec{w}$. To find the rotation angle ϑ, it suffices to calculate the angular velocity of this rotation, which is uniform when the angle α that defines the quaternion q changes uniformly.

Replacing $\cos\alpha$ and $\sin\alpha$ by their Taylor series at $\alpha = 0$, and calculating only the linear terms in α, considered as a small quantity, for a purely imaginary quaternion $\vec{h}$ orthogonal to $\vec{w}$ in $\mathbb{R}^3$, we get its first order shifting by a small rotation g, with an error of order α^2:

$$\begin{aligned}
q\vec{h}q^{-1} &= (1 + \alpha\vec{w})h(1 - \alpha\vec{w}) = (\vec{h} + \alpha[\vec{w}, \vec{h}])(1 - \alpha\vec{w}) \\
&= \vec{h} + \alpha[\vec{w}, \vec{h}] - \alpha[\vec{h}, \vec{w}] \\
&= \vec{h} + 2\alpha[\vec{w}, \vec{h}]\,.
\end{aligned}$$

This expression shows that the vector $\vec{h}$ rotates around the axis $\vec{w}$ with angular velocity $2\vec{w}$, and hence $\vartheta = 2\alpha + o(\alpha)$. The choice of $\alpha = \vartheta/2$ in the definition of the quaternion q would produce a rotation g with angle $\vartheta + o(\vartheta)$, for small values of ϑ. This implies the exact formula (3) for any ϑ, since the map sending q to g is a group homomorphism and then ϑ is a linear function of α.

Hamilton's theory describing the rotations of the 3-space by the quaternions $q \in \mathbb{S}^3 \subset \mathbb{H}$ is thus proved.

We have constructed the two-fold covering $\mathbb{S}^3 \to \mathrm{SO}(3)$, which is the group homomorphism sending the quaternion q of norm 1 to the rotation g. Topologically it is just the universal covering of the manifold SO(3) by its spin-group $\mathrm{Spin}(3) = \mathbb{S}^3$.

Earlier, we had proved only the homeomorphism $\mathrm{SO}(3) \sim \mathbb{R}\mathrm{P}^3$. Now, the Hamilton covering provides us the following corollary.

Corollary. *The manifold* SO(3) *of the special orthogonal matrices of order* 3 *is diffeomorphic to the real projective* 3*-space:* $\mathrm{SO}(3) \approx \mathbb{R}\mathrm{P}^3 = \mathbb{S}^3/\{\pm 1\}$.

Problem. Prove that the group $\mathbb{S}^3 = \mathrm{Spin}(3)$ coincides with the group SU(2) of the 2×2 unitary matrices of determinant 1.

Hint. A quaternion may be considered as a pair of complex numbers.

3.12 Dirac's experiment on spherical braids

The 2-fold spin covering is very useful in modern physics. The quantum physicists proved experimentally that the "electron quantum rotations" should be described rather by their "spins" (quaternions) than by the rotations g of the macroscopic solids. One should imagine that the interior quantum structure of the electron is better described by the corresponding quaternion q than by the classical rotation g, which makes irrelevant whether the quaternion is q or $-q$. Neighbouring electrons admitting equal rotations g dislike to have equal spins q, "the Pauli exclusion principle".

Creating this spin theory of the electron, P. Dirac decided to show experimentally the relevant topological facts to the physicists. He had in mind the finiteness of the fundamental group $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$, but the group of rotations SO(3) is not an object easy to demonstrate. So Dirac invented a corollary that is easier to observe.

His example comes from braid theory. The braid group of n strings $\mathrm{Br}(n)$, studied with details in Ch. 5, p. **??**-144, is the fundamental group of the manifold formed by the sets of n distinct points of the plane $\mathbb{R}^2$. Its braids can be represented by the graphs of the motions of n points in $\mathbb{R}^3$. Such a graph consists of n disjoint paths in $\mathbb{R}^3$ that join n points of the plane $z = 0$ to the n points of the plane $z = -1$ with the same coordinates x and y, and that are transverse to the planes $z = \text{const}$ – Fig. 3.22.

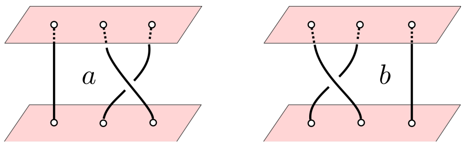


Figure 3.22: Two braids of 3 strings.

An element of $\mathrm{Br}(n)$ is a homotopy class of such representatives, for a fixed set of end-points. The product of two strings a and b is represented by the attachment of b below a (between $z = -1$ and $z = -2$).

The braid groups $\mathrm{Br}(n)$ have no elements of finite order (see Chapter 5). If some braid γ is non trivial, then its repetitions γ^2, γ^3, $\dots$ are also non trivial.

Dirac suggested to consider, together with the ordinary braids, the braids on arbitrary surfaces M^2. The group $\mathrm{Br}(n, M^2)$ is defined as the fundamental group of the manifold of the unordered configurations of n different points of the surface M^2:

$$\mathrm{Br}(n, M^2) = \pi_1(\mathrm{Conf}_n(M^2))\,.$$

Consider for instance the case of the braids on the sphere, $M^2 = \mathbb{S}^2$. Such a braid can be realised physically by connecting two concentric spheres by n strings whose projections to the radial direction are injective (Fig. 3.23).

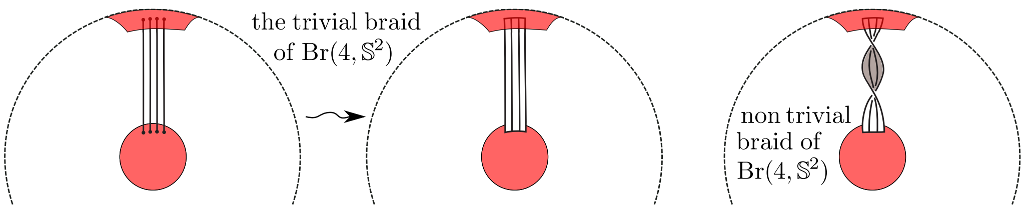


Figure 3.23: Spherical braids of 4 strings.

In Fig. 3.23-middle, we have glued the strings of the trivial braid to a ribbon. To get the non trivial braid of Fig. 3.23-right, we wind the ribbon extremity by the angle 2π.

The order 2 of the fundamental group of the group of the 3-space rotations (providing the spins) is also the origin of the strange existence of the braids of finite order on the sphere.

Dirac shown experimentally a second order braid of four strings in $\mathrm{Br}(4, \mathbb{S}^2)$. The experimental design consisted of three concentric spheres connected by four ropes intersecting each of the three spheres at the points of its intersections with four radial rays. The four ropes were not straight, but making turns one around the other, like in the case of ordinary braids.

The braid connecting the larger sphere to the middle one was topologically the same as the braid connecting the middle sphere to the smallest one.

Dirac destroyed the middle sphere. Then the ropes, free from the obstruction of this middle sphere, formed a trivial braid: The four ropes, connecting the larger sphere to the smallest one, arrived to the state of purely radial straight lines by a homotopy, and the physicists were able to observe it.

Now, we shall explain the mathematical side of Dirac's experiment.

Consider a closed curve in $\mathrm{SO}(3)$ as a continuous family of orthonormal frames in $\mathbb{R}^3$. We associate to each frame $\{u, v, w\}$ four points on the sphere $\mathbb{S}^2 \subset \mathbb{R}^3$ of radius 1: The end-points of the vectors $\{u, v, w, (u+v+w)/\sqrt{3}\}$.

These four points are different and define hence a configuration belonging to the space $\mathrm{Conf}_4(\mathbb{S}^2)$. Any closed curve γ in $\mathrm{SO}(3)$ generates in this way a spherical braid b_γ belonging to $\mathrm{Br}(4, \mathbb{S}^2)$.

Choosing a particular curve γ in $\mathrm{SO}(3)$, Dirac obtained his second order spherical braid. Indeed, the product of the spherical braid b_γ with itself is trivial in $\mathrm{Br}(4, \mathbb{S}^2)$ because the homotopy of the doubled curve $\gamma \cdot \gamma$ to the trivial curve $*$ in $\mathrm{SO}(3)$ induces a homotopy of the doubled braid to the trivial one in $\mathrm{Br}(4, \mathbb{S}^2)$.

It remains to prove that the initial braid on the sphere was a non trivial element of the spherical braid group $\mathrm{Br}(4, \mathbb{S}^2)$.

We shall do it by geometric considerations on elliptic curves. Namely, associate to the configuration of four points on the sphere the torus surface of the elliptic curve obtained as the double covering of the sphere, ramified at these four points.

The topological construction of the torus is the following: one connects the ramification points z_1 and z_2 by one simple cut and the ramification points z_3 and z_4 on the sphere by another disjoint simple cut to create the two sheets of the covering. To obtain the torus surface of the elliptic curve, one glues these two sheets along the cutting segments – Fig. 3.24.

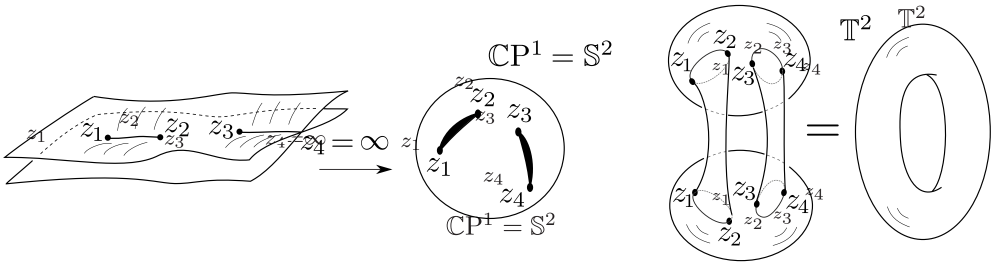


Figure 3.24: Elliptic curve construction.

The homotopy class of a loop in the torus is a linear combination with integral coefficients, $au + bv$, of the homotopy classes u and v of the basic loops of the torus. Identifying (the classes of) these loops with the respective vectors $(1, 0)$ and $(0, 1)$, the monodromy representation of the spherical braid is a 2×2 matrix that describes the homeomorphism action on the basic loops u and v. The spherical braid is non trivial, whenever its monodromy matrix is not the identity.

In order to calculate this matrix, observe that for each string of our spherical braid the initial point and the end point coincide because for a loop in the space of frames $\{u, v, w\}$, the initial ordered frame coincides with the final ordered frame. The most simple non trivial braid of such type should be provided by a 2π rotation around the diagonal vector.

We can get such a rotation using the generating braids a and b of Fig. 3.22. The reader can verify that the braid aba produces half of the 2π rotation we need, and hence $(aba)^2$ provides our 2π rotation. Using the respective matrices $\left(\begin{smallmatrix} 1 & 0 \\ -1 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ of the generating braids a and b (whose explicit calculation is given in Sect. 5.9, p. 161-164) we calculate the monodromy matrix $\left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ of our spherical braid $(aba)^2$. It is not the identity monodromy and its square is the identity, as it must be for a second order element of the fundamental group represented by the monodromy.

In this way Dirac's experiment relates the second order elements of the fundamental group of the group of rotations (and the spins covering the rotations) to the second order elements of the braid group of four strings on the sphere.

More convincing than the above mathematical arguments is to see a concrete spherical braid whose square is the trivial braid. In Fig. 3.25, we show that the non trivial braid of Fig. 3.23-right is an order two element of $\mathrm{Br}(4, \mathbb{S}^2)$: its square is the trivial braid.

In fact, Fig. 3.25 shows also that for any $n \geqslant 3$ there is a non trivial element of order two: we have just to glue n strings to the ribbon of Figures 3.23 and 3.25.

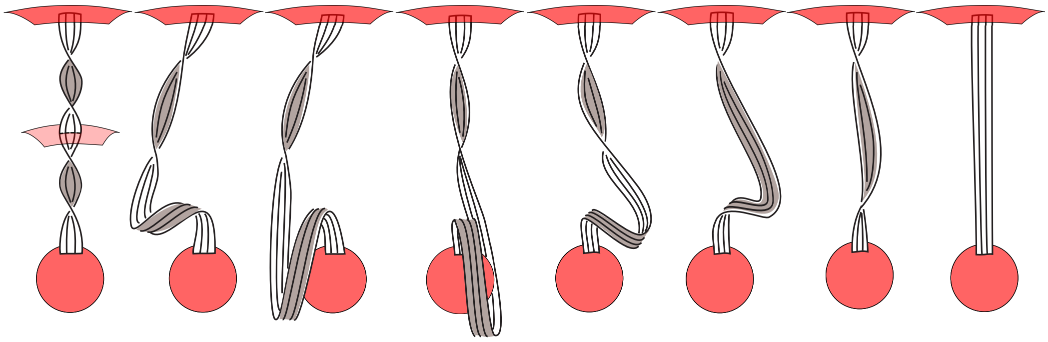


Figure 3.25: The square of an order 2 element of the braid group $\mathrm{Br}(n, \mathbb{S}^2)$.

It is interesting that the great physicist Dirac always insisted on the following clever use of mathematics in physical theories:

"*I learnt to distrust all physical concepts as the basis for a theory. Instead one should put one's trust in a mathematical scheme, even if the scheme does not appear at first sight to be connected with physics. One should concentrate on getting an interesting mathematics*". [We quote his words as reported in P. Masani "Norbert Wiener", Birkhäuser, 1970, p. 6].

Dirac's idea on the physical concepts of the previous generations was that this is a polite name for the obsolete traditional prejudices.

The celebrated "Wigner principle" claims that the observable effectiveness of the mathematical models in so many applied problems of physics and of other sciences has no reasonable explanation. As I. Gelfand formulated it, the only comparable phenomenon is the equally unreasonable non effectiveness of mathematics in biology.

Consider the action of the product $\mathbb{S}^3 \times \mathbb{S}^3$ of two copies of the 3-sphere of quaternions of norm 1 (in Euclidean space $\mathbb{R}^4 = \mathbb{H}$) on the quaternions $h \in \mathbb{H}$, sending h to $Ah = q'hq^{-1}$. Since the quaternions q' and q are of norm 1, the linear operator $A : \mathbb{R}^4 \to \mathbb{R}^4$ is orthogonal: $A \in \mathrm{SO}(4)$.

Problem. Prove that *every rotation of the 4-space may be described in this way, and that the map sending the pair* (q, q') *to* A *is the two-fold universal covering and is a homomorphism*[*] *of groups,* $\mathbb{S}^3 \times \mathbb{S}^3 \to \mathrm{SO}(4)$, *interpreting*

[*] The multiplication from the right by q^{-1} rather than by q is needed for this goal.

the group Spin(4) *as the direct product* $\mathbb{S}^3 \times \mathbb{S}^3$ *of two copies of the group of quaternions of length* 1.

3.13 Relative homotopy groups

Let Y be a submanifold of a manifold X, consider the k-cube as the product $I^k = I^{k-1} \times I$ and write I_0^{k-1} for its face $I^{k-1} \times \{0\}$.

Definition. The kth *relative homotopy group* $\pi_k(X, Y, *)$ of the pair (X, Y) with base-point $* \in Y$ consists of the homotopy classes of the maps of triples

$$f : (I^k, \partial I^k, I_0^{k-1}) \to (X, Y, *)\,, \qquad k > 1\,,$$

where $f(\partial I^k) \subset Y,\ f(I_0^{k-1}) = * \in Y$.

Remark. Write J^{k-1} for the closure of the complement to I_0^{k-1} in ∂I^k, that is, $J^{k-1} := \overline{\partial I^k \setminus I_0^{k-1}}$. Since J^{k-1} is homeomorphic to the standard cube I^{k-1} and the boundaries of the face I_0^{k-1} and of J^{k-1} coincide, the restriction of f to the "cube" J^{k-1}, is a $(k-1)$-spheroid at $* \in Y$, $f_{|J^{k-1}} : J^{k-1} \to Y$.

The set of classes has a group structure defined in a similar way to that of the absolute homotopy group. For $k > 2$, the relative homotopy groups are Abelian.

The following three natural homomorphisms appear in the study of relative homotopy groups:

$$j_* : \pi_n(X, *) \to \pi_n(X, Y, *),$$

$$\partial : \pi_n(X, Y, *) \to \pi_{n-1}(Y, *),$$

$$i_* : \pi_n(Y, *) \to \pi_n(X, *).$$

With these homomorphisms the relative homotopy groups form a natural long sequence, called *homotopy sequence of the pair* (X, Y):

$$\cdots \xrightarrow{\partial} \pi_k(Y, *) \xrightarrow{i_*} \pi_k(X, *) \xrightarrow{j_*} \pi_k(X, Y, *) \xrightarrow{\partial} \pi_{k-1}(Y, *) \xrightarrow{i_*} \cdots\,.$$

To define j_*, one uses the fact that any map $f : (I^k, \partial I^k) \to (X, *)$ can be considered as a map of triples $h : (I^k, \partial I^k, I_0^{k-1}) \to (X, Y, *)$, since $* \in Y$.

To define the homomorphism ∂ one uses the restrictions of the maps $f : (I^k, \partial I^k, I_0^{k-1}) \to (X, Y, *)$ to ∂I^k, $f_{|\partial I^k} : (\partial I^k, I_0^{k-1}) \to (Y, *)$.

The inclusion homomorphism i_* is induced from the inclusion $Y \subset X$.

Theorem 13. *The homotopy sequence of any pair (X, Y) is exact.*

Problem. Prove Theorem 11.

Hint. You have to prove the three equalities

$$\operatorname{Im} i_* = \operatorname{Ker} j_* , \quad \operatorname{Im} j_* = \operatorname{Ker} \partial , \quad \operatorname{Im} \partial = \operatorname{Ker} i_* .$$

To prove the inclusions $\operatorname{Im} i_* \subset \operatorname{Ker} j_*$, $\operatorname{Im} j_* \subset \operatorname{Ker} \partial$, $\operatorname{Im} \partial \subset \operatorname{Ker} i_*$, show that each of the following respective compositions is the zero homomorphism: i) $j_* \circ i_* = 0$, ii) $\partial \circ j_* = 0$, iii) $i_* \circ \partial = 0$.
i) The composition $j_* \circ i_*$ is zero because every map $(I^n, \partial I^n, I_0^{n-1}) \to (Y, *, *)$ represents zero in $\pi_n(X, Y, *)$.
ii) $\partial \circ j_*$ is zero because the restriction of a map $(I^n, \partial I^n, I_0^{n-1}) \to (X, *, *)$ to J^{n-1} has $*$ as its image, and hence represents zero in $\pi_{n-1}(Y, *, *)$.
iii) The composition $i_* \circ \partial$ is zero because the restriction of any map $f : (I^k, \partial I^k, I_0^{k-1}) \to (X, Y, *)$ to J^{k-1} is homotopic to a constant map via f itself, by a homotopy whose restriction to ∂J^{k-1} is independent of t.

We leave to the reader the pleasure to prove the remaining inclusions: $\operatorname{Im} i_* \supset \operatorname{Ker} j_*$, $\operatorname{Im} j_* \supset \operatorname{Ker} \partial$, $\operatorname{Im} \partial \supset \operatorname{Ker} i_*$.

This exact sequence is used almost only in the case where the submanifold $Y = F$ is a fibre of a fibration $p : X \to B$. In this case the isomorphism $\pi_k(X, F) \simeq \pi_k(B)$ holds, obtaining the exact sequence that we have already considered.

3.13.1 On Quotient Homotopy Groups

Given a submanifold Y of a manifold X, it is also natural to investigate the *quotient homotopy groups* $\pi_k(X/Y, *)$, which are defined as the ordinary homotopy groups of the space X/Y obtained from X by the gluing of the whole subset Y to one point $*$. We suppose here that the distinguished point $*$ of X belongs to Y.

Thus the paths that define the relative fundamental group $\pi_1(X/Y, *)$ are not closed curves in X that start and end at the point $* \in X$: They are the paths in X connecting arbitrary points of Y.

Example. If $X = \mathbb{S}^1$ is a circle, and $Y = \mathbb{S}^0$ its intersection with a diameter, then the space X/Y is the curve "∞" (equipped with the singular point $*$), and thus

$$\pi_1(X/Y, *) = F_2$$

is the free group generated by the two closed circles of "∞".

Remark. The quotient homotopy groups form the natural sequence

$$
\begin{tikzcd}
\cdots \arrow[r] & \pi_k(Y,*) \arrow[r, "i_*"] & \pi_k(X,*) \arrow[r, "j_*"] & \pi_k(X,Y,*) \arrow[r, "\partial"] \arrow[dr] & \pi_{k-1}(Y,*) \arrow[r] & \cdots \\
& & & & \pi_k(X/Y) \arrow[u, "?"'] &
\end{tikzcd}
.
$$

This horizontal sequence can be in some cases continued by the map labeled here with " ? ", which has not always a natural definition. Such continuation could be some times even exact or semi-exact. We leave to the reader the pleasure to investigate such cases and to write by himself all the details (for example, in the case where Y is contractible). Both, the fact that the above sequence is not always continued (because the map " ? " is not always defined) and its possible non exactness, when it can be continued, reflect the differences between homotopy theory and homology theory (studied in Ch. 11) where the corresponding maps are well defined and the exactness holds. The main difficulty here is that the map $X \to X/Y$ may have non-liftable homotopies.

Problem. Calculate the homotopy groups of the curve "∞".

Answer. For $k \geqslant 2$ the isomorphism $\pi_k(\text{“}\infty\text{”}) \approx \pi_k(\mathbb{S}^1)$ takes place, since $\pi_k(\mathbb{S}^0) = \pi_{k-1}(\mathbb{S}^0) = 0$, and hence all these groups are trivial.

For $k = 1$ the part of sequence

$$\pi_1(\mathbb{S}^0) \to \pi_1(\mathbb{S}^1) \to \pi_1(\text{“}\infty\text{”}) \to \pi_0(\mathbb{S}^0) \to \pi_0(\mathbb{S}^1)\,,$$

is not exact because $\pi_1(\text{“}\infty\text{”})$ is the free group with two generators.

Problem. Calculate the homotopy groups of the complement to two points of the plane.

Solution. The groups $\pi_k(\mathbb{R}^2 \smallsetminus \mathbb{S}^0)$ are trivial for $k \geqslant 2$, being naturally isomorphic to those of the curve "∞".

Indeed, this curve can be included into the complement to two points, $i : \text{“}\infty\text{”} \to \mathbb{R}^2 \smallsetminus \mathbb{S}^0$, in such a way that the whole complement can be retracted to the curve "∞" by a homotopy $\{g_t\}$ that preserves every point of this curve at its place, starts from the identity map g_0 of the complement to itself and ends at the projection $g_1 : (\mathbb{R}^2 \smallsetminus \mathbb{S}^0) \to \text{“}\infty\text{”}$. This deformation retract provides the isomorphism $g_{1*} : \pi_k(\mathbb{R}^2 \smallsetminus \mathbb{S}^0) \to \pi_k(\text{“}\infty\text{”})$.

Problem. Calculate the homotopy groups π_k, $k \geqslant 2$, of the complement to n points of the plane.

Hint. Show that the complement to n points of the plane can be retracted to a "lemniscate" curve consisting of n circles joined at one common point. This implies that $\pi_k(\mathbb{R}^2 \smallsetminus \{n \text{ points}\}) = \pi_k(\mathbb{S}^1/Y)$, where Y is a set of n different points of $\mathbb{S}^1$. Then, prove that $\pi_k(\mathbb{S}^1/Y, *) = \pi_k(\mathbb{S}^1, *) = 0$, for $k \geqslant 2$.

Problem. Calculate the homotopy group π_2 of the complement to the union of a straight line and a point disjoint to this line in Euclidean space $\mathbb{R}^3$.

Hint. Reduce this homotopy group to the form $\pi_2(B)$, where the space B consists of a sphere $\mathbb{S}^2$ joined to a circle $\mathbb{S}^1$ at a single common point $*$.

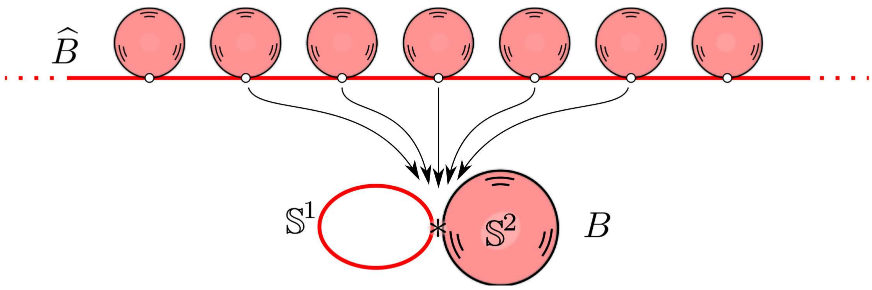


Figure 3.26: The universal covering space $\widehat{B}$ of the union B of the sphere $\mathbb{S}^2$ with a circle $\mathbb{S}^1$ that intersects it at one point $*$.

The universal covering space $\widehat{B}$ of B consists of an infinite number of disjoint spheres $\mathbb{S}^2$ and of the real line $\mathbb{R}$, such that the jth sphere has with the line one common point: $j \in \mathbb{Z}$ (Fig. 3.26).

The homotopy exact sequence of the fibration $\widehat{B} \to B$ provides the isomorphism $\pi_2(B) \approx \pi_2(\widehat{B})$. Thus the group $\pi_2(B)$ has an infinite number of independent elements φ_j, corresponding to the spheroid covering the jth sphere of $\widehat{B}$.

We shall prove below that $\pi_2(\mathbb{S}^2) \approx \mathbb{Z}$, and hence that the group $\pi_2(\widehat{B}) \approx \pi_2(B)$ is the infinite product $\mathbb{Z}^\infty$, generated by the independent spheroids φ_j.

Problem. Find the homotopy groups π_k, $k \geqslant 2$, of the space $\mathrm{Conf}_n(\mathbb{R}^2)$ of unordered configurations of n non-coincident points in the plane.

Solution. Cover the space $\mathrm{Conf}_n(\mathbb{R}^2)$ by the space $M_n = \widetilde{\mathrm{Conf}_n}(\mathbb{R}^2)$ of the ordered configurations. The covering is a fibration whose fibre consists of the

$n!$ possible orderings of a given configuration. Thus the homotopy groups π_k $(k \geqslant 2)$ of the space of ordered configurations and of the space of the unordered ones are isomorphic.

In order to compute the homotopy groups of the space M_n of ordered configurations of n points, we map this space to M_{n-1}, associating to each ordered configuration its first $n-1$ points. This map is a fibration over M_{n-1} whose fibre F_{n-1} is the complement to some $n-1$ points of the plane. The homotopy groups π_k of such fibres are trivial for $k \geqslant 2$, as we have seen above using the “lemniscate” type curves.

Thus the short exact sequences of this fibration,

$$\pi_k(F_{n-1}) \to \pi_k(M_n) \to \pi_k(M_{n-1}) \to \pi_{k-1}(F_{n-1})\,,$$

provide the isomorphisms between the groups π_k of the spaces M_n and M_{n-1} only for $k \geqslant 3$ because $\pi_1(F_{n-1}) \neq 0$. Since $\pi_k(M_1) = 0$ for $k \geqslant 1$ (because $M_1 = \mathbb{R}^2$), this sequence of isomorphisms provides the (triviality) answer:

$$\pi_k(\mathrm{Conf}_n(\mathbb{R}^2)) \approx \pi_k(M_n) \approx 0\,, \quad \text{for} \;\; k \geqslant 3\,.$$

For $k = 2$ we use the equality $\pi_2(M_1) = 0$ in the short sequence

$$0 = \pi_2(F_1) \to \pi_2(M_2) \to \pi_2(M_1) = 0\,,$$

which together with the short sequences

$$0 = \pi_2(F_{n-1}) \to \pi_2(M_n) \to \pi_2(M_{n-1})\,, \quad \text{for } n > 2\,,$$

imply the triviality:

$$\pi_k(\mathrm{Conf}_n(\mathbb{R}^2)) \approx \pi_k(M_n) \approx 0\,, \quad \text{for} \;\; k = 2\,.$$

The fundamental group $\pi_1(\mathrm{Conf}_n(\mathbb{R}^2))$ deserves a special study. In Chapter 5 (p. 138), we show that $\pi_1(\mathrm{Conf}_n(\mathbb{R}^2))$ is naturally isomorphic to the braid group of n strings, which is calculated as the fundamental group of the complement to a complex algebraic hypersurface in $\mathbb{C}^{n-1}$.

Problem. Calculate the homotopy groups π_k, $k \geqslant 2$, of the surface M^2 of genus $g > 0$.

Hint. The universal covering of such a surface M^2 is diffeomorphic to the plane $\mathbb{R}^2$. The homotopy exact sequence of this covering provides the answer:

$$\pi_k(M^2) = 0\,, \qquad \text{for} \;\; k \geqslant 2\,.$$

Problem. Calculate the fundamental group of the group $\mathrm{SL}(n, \mathbb{R})$ of the real matrices of order n with determinant 1.

Solution. The space of matrices of positive determinant and the space of matrices of determinant 1 are related by a natural fibration with 1-dimensional contractible fibres and hence their homotopy groups are isomorphic.

To solve the problem, we have to prove that these homotopy groups are also isomorphic to the homotopy groups of the group of orthogonal matrices of determinant 1:

$$\pi_k(\mathrm{SL}(n, \mathbb{R})) \approx \pi_k(\mathrm{SO}(n, \mathbb{R}))\,.$$

We identify the manifolds of these groups of matrices with some manifolds of frames. Namely, we associate each element g of the group of matrices with the image of a chosen standard n-frame under g.

Let E be the manifold formed by the frames that orient Euclidean space $\mathbb{R}^n$ in a fixed "positive" way, and let B be the manifold of such orthonormal frames in the same space.

The standard orthonormalisation algorithm associates to every positive frame $(e_1, \dots, e_n)$ its orthonormalised version $(\widetilde{e}_1, \dots, \widetilde{e}_n)$, consisting of orthogonal vectors of length 1. The first vector, $\widetilde{e}_1$, is the normalised version of e_1, being positively proportional to it. The second, $\widetilde{e}_2$, is the normalised projection of e_2 on the orthocomplementary direction to e_1 in the plane generated by e_1 and e_2; and so on: $\widetilde{e}_k$ belongs to the k-dimensional plane generated by $(e_1, \dots, e_k)$.

The resulting orthogonalisation map $p : E \to B$ is a fibration. Its fibre, which consists of the frames $(e_1, \dots, e_n)$ with coinciding normalisations $(\widetilde{e}_1, \dots, \widetilde{e}_n)$, is contractible. Say, for $n = 3$, knowing $(\widetilde{e}_1, \widetilde{e}_2, \widetilde{e}_3)$, we can take as e_1 an arbitrary positive multiple of $\widetilde{e}_1$, as e_2 an arbitrary vector in the half-plane generated by $\widetilde{e}_1$ and the positive multiples of $\widetilde{e}_2$, and as e_3 an arbitrary vector of the half-space containing $\widetilde{e}_3$, separated by the preceding plane. The manifold of the arbitrary choices is thus contractible.

The homotopy exact sequence of the fibration implies the promised isomorphisms:

$$\pi_k(\mathrm{GL}^+(n, \mathbb{R})) \approx \pi_k(\mathrm{SL}(n, \mathbb{R})) \approx \pi_k(\mathrm{SO}(n, \mathbb{R}))\,.$$

Corollary. *The fundamental group of the manifold of real matrices of order $n \geqslant 3$ of determinant 1 consists of two elements, and it is $\mathbb{Z}$ for $n = 2$:*

$$\pi_1(\mathrm{SL}(n, \mathbb{R})) \approx \mathbb{Z}_2\,, \qquad \textit{for} \;\; n \geqslant 3\,.$$

3.14 On Bott periodicity theorems

Similarly to the difficult homotopy groups of the spheres, most part of the homotopy groups of the groups GL^+, SL, or of their orthogonal subgroups SO, have not been computed, even for SO(3).

However, a peculiar "Bott periodicity" theorem states that the *stable homotopy groups* $(\pi_k(\mathrm{SO}(n,\mathbb{R})),\, n \to \infty)$ are periodic in k, with period 8:

$$\pi_k(\mathrm{SL}(\infty)) \approx \pi_k(\mathrm{SO}(\infty)) \approx \pi_{k+8}(\mathrm{SO}(\infty))\,.$$

These eight groups have been already computed, similarly to the above computation of

$$\pi_1(\mathrm{SL}(\infty)) \approx \pi_1(\mathrm{SO}(\infty)) = \mathbb{Z}_2\,.$$

Theorem 14 (Bott)**.** *The eight stable homotopy groups* $\pi_k(\mathrm{SO}(\infty))$ *are*:

$k \pmod 8$	0	1	2	3	4	5	6	7
$\pi_k(\mathrm{SO}(\infty))$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$

Remark. Combining this result with Theorem 10 (p. 82), we obtain all the homotopy groups $\pi_k(\mathrm{SO}(n))$ in the so-called "stable range" $k \leqslant n-2$. Besides that, these groups are known only for few slightly larger values of k.

On the ideas of the proof of the Bott periodicity Theorem. The main geometric trick is an imitation of the construction of the equator $\mathbb{S}^{n-1}$ of the sphere $\mathbb{S}^n$: One considers the geodesics connecting two fixed points, say, the North and the South poles of the sphere, equipped with the standard metric of constant positive curvature. These are meridians, and they are naturally parametrised by the points of the equator $\mathbb{S}^{n-1}$.

In a similar way, Bott constructed an "equator" submanifold consisting of the minimal geodesics from e to $-e$ on $\mathrm{O}(2^4 m)$. Repeating this operation by constructing the "equator of the equator" and so on, he obtained eight interesting classical manifolds, arriving after eight steps to a manifold that is "approximately" the smaller orthogonal group of matrices $\mathrm{O}(m)$.

The word "approximately" here is due to the presence of some additional cells of higher dimensions, which are unable to change the homotopy groups of smaller dimension that we study.

We leave to the reader the pleasure to study this periodicity theorem and the details of the proof in the nice book "Morse Theory" of J. Milnor [102].

Unitary Groups. We shall use again the homotopy exact sequences.

Exercise. Calculate the homotopy groups $\pi_3(\mathrm{U}(n))$.

Solution. For $n = 2$ the homotopy exact sequence associated to the "first vector" fibration $\mathrm{U}(2) \to \mathbb{S}^3$ (with fibre $\mathbb{S}^1$) contains the fragment

$$(\pi_3(\mathbb{S}^1) = 0) \to \pi_3(\mathrm{U}(2)) \to \pi_3(\mathbb{S}^3) \to (\pi_2(\mathbb{S}^1) = 0)\,.$$

Hence the equality $\pi_n(\mathbb{S}^n) = \mathbb{Z}$ (see p.362) provides the answer $\pi_3(\mathrm{U}(2)) = \pi_3(\mathbb{S}^3) = \mathbb{Z}$.

Similarly, for $n \geqslant 2$ the first vector fibration $p : \mathrm{U}(n+1) \to \mathbb{S}^{2n+1}$, with fibre $\mathrm{U}(n)$, provides the fragment

$$(\pi_4(\mathbb{S}^{2n+1}) = 0) \to \pi_3(\mathrm{U}(n)) \to \pi_3(\mathrm{U}(n+1)) \to (\pi_3(\mathbb{S}^{2n+1}) = 0),$$

and whence the answers $\pi_3(\mathrm{U}(n+1)) \approx \pi_3(\mathrm{U}(n)) \approx \ldots \approx \pi_3(\mathrm{U}(2)) \approx \mathbb{Z}$.

Problem. Prove the stabilisation (and the stable range $k \leqslant 2n-1$) of the homotopy groups of the manifolds $U(n)$ of unitary matrices

$$\pi_k(\mathrm{U}(n)) \approx \pi_k(\mathrm{U}(n+1)) \approx \ldots = \pi_k(\mathrm{U}(\infty))\,.$$

(In particular we have $\pi_1(\mathrm{U}(n)) \approx \pi_1(\mathrm{U}(1)) \approx \pi_1(\mathbb{S}^1) = \mathbb{Z}$.)

Hint. Similarly to the case of the special orthogonal group (p. 81), the "first vector map" of the frames in $\mathbb{C}^{n+1}$ provides a natural fibration

$$p : \mathrm{U}(n+1) \to \mathbb{S}^{2n+1}\,,$$

whose fibre is the unitary group $\mathrm{U}(n)$ (which represents the remaining part of the Hermitian frame). The fragments of its homotopy exact sequence and the fact that $\pi_\ell(\mathbb{S}^m) = 0$ if $0 < \ell < m$ provide the stabilisation isomorphisms

$$\pi_k(\mathrm{U}(n)) \approx \pi_k(\mathrm{U}(n+1))\,,$$

for $k+1 < 2n+1$, bringing the stable range $k \leqslant 2n-1$.

The periodicity of the stable groups in the complex case is simpler, since the period 8 is reduced to 2:

$$\pi_k(\mathrm{GL}(\infty, \mathbb{C})) \approx \pi_k(\mathrm{U}(\infty)) \approx \pi_{k+2}(\mathrm{U}(\infty))\,,$$

where $\mathrm{GL}(n, \mathbb{C})$ denotes the *general linear group*, which is the group of the non degenerate complex linear operators in $\mathbb{C}^n$ or of the complex matrices of order n of non-zero determinant.

The stable homotopy groups of the unitary groups are now easily obtained, since we know that $\pi_0(\mathrm{U}(1)) = 0$ and $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$:

$k \pmod 2$	0	1
$\pi_k(\mathrm{U}(\infty))$	0	$\mathbb{Z}$

.

3.15 Symplectic Group, Lagrange Grassmannian and Maslov Index

Symplectic space. The *standard symplectic space* is the space $\mathbb{R}^{2n}$ with "*Darboux coordinates*" $(p_1, q_1, \ldots, p_n, q_n)$, equipped with the bilinear form, called *standard symplectic structure*,

$$\omega(\xi, \eta) = \sum_{k=1}^{n} (p_k(\xi) q_k(\eta) - p_k(\eta) q_k(\xi)) .$$

A linear operator $A : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is said to be *symplectic* if it preserves the bilinear form ω :

$$\omega(A\xi, A\eta) = \omega(\xi, \eta) , \qquad \text{for any } \ \xi, \eta \in \mathbb{R}^{2n} .$$

Symplectic group. The subgroup $\mathrm{Sp}(n, \mathbb{R}) \subset \mathrm{GL}(2n, \mathbb{R})$ formed by all the symplectic linear operators on $\mathbb{R}^{2n}$ is called *symplectic group*.

Problem. Calculate the dimension and the fundamental group of the smooth manifold $\mathrm{Sp}(n, \mathbb{R})$.

Answer. $\dim(\mathrm{Sp}(n, \mathbb{R})) = n(2n + 1)$, $\pi_1(\mathrm{Sp}(n, \mathbb{R})) \approx \mathbb{Z}$.

The stable homotopy groups $(\pi_k(\mathrm{Sp}(n, \mathbb{R})),\ n \to \infty)$ have also period 8 in k and their stable range is $2 \leqslant k \leqslant 4n + 1$.

Theorem 15. *The eight stable homotopy groups* $\pi_k(\mathrm{Sp}(\infty))$ *are*:

$k \pmod 8$	0	1	2	3	4	5	6	7
$\pi_k(\mathrm{Sp}(\infty))$	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$

Lagrange subspaces. An n-dimensional subspace of the symplectic vector space $\mathbb{R}^{2n}$ is called *Lagrangian* if $\omega(\xi, \eta) = 0$ for any two vectors ξ, η of this subspace. Hence a Lagrangian subspace is its own symplectic orthocomplement (according to F. Klein, because of that Sophus Lie had initially called them "mad subspaces").

The set of all Lagrangian (non oriented) subspaces of a fixed space $\mathbb{R}^{2n}$ is a smooth manifold $\Lambda(n)$, called the *Lagrangian Grassmannian*.

Problem. Calculate the dimension and the fundamental group of the Lagrangian Grassmannian manifold $\Lambda(n)$.

Solution. Consider the Hermitian product $\langle z, w\rangle_{\mathbb{C}} = \sum_{\kappa=1}^{n} z_k \overline{w}_k$ in $\mathbb{C}^n$. The real Euclidean scalar product $\langle z, w\rangle_{\mathbb{R}} = \sum_{\kappa=1}^{n}(p_k(z)p_k(w) + q_k(z)q_k(w))$ in $\mathbb{R}^{2n}$ and the real symplectic product ω in $\mathbb{R}^{2n}$ are related by the obvious formula:

$$\langle z, w\rangle_{\mathbb{C}} = \langle z, w\rangle_{\mathbb{R}} - i\omega(z, w)\,. \tag{1}$$

A real linear operator in $\mathbb{R}^{2n}$ is said to be *complex* if it preserves the multiplication by i,

$$A(iz) = iAz\,, \qquad \text{for any} \;\; z\,.$$

Such operators form the complex general linear group $\mathrm{GL}(n, \mathbb{C})$. These operators are orthogonal (that is, belonging to $O(2n)$) if they preserve the scalar product

$$\langle Az, Aw\rangle_{\mathbb{R}} = \langle z, w\rangle_{\mathbb{R}} \qquad (\text{for any} \;\; z, w\,)$$

and are symplectic if they preserve the symplectic form ω.

We get three subgroups of the group of general linear operators $\mathrm{GL}(2n, \mathbb{R})$. Namely $\mathrm{GL}(n, \mathbb{C})$, $\mathrm{O}(2n)$ and $\mathrm{Sp}(n, \mathbb{R})$. Formula (1) implies the important

Theorem 16. *The three pairwise intersections of these three subgroups coincide.*

That is, if an operator preserves the complex structure and the lengths, then it is symplectic, and if it is symplectic and orthogonal, then it preserves the Hermitian structure.

It is useful to remark the (obvious) relation between the three structures of formula (1):

$$\omega(\xi, \eta) = -\langle i\xi, \eta\rangle_{\mathbb{R}}\,.$$

The intersection of the three subgroups is called the *unitary group* $\mathrm{U}(n)$ (Fig. 3.27).

The image of the real subspace $\mathbb{R}^n$ (where $q = 0$) under a unitary operator $g \in \mathrm{U}(n)$ is a Lagrangian subspace in $\mathbb{R}^{2n}$, since the initial real subspace is Lagrangian and g preserves the symplectic form ω.

We get in this way a map $\mathrm{U}(n) \to \Lambda(n)$, sending g to the subspace $g\mathbb{R}^n$. This map is a smooth fibration over the whole Lagrangian Grassmannian manifold, whose fibre over the initial point $\mathbb{R}^n$ of $\Lambda(n)$ consists of the unitary operators preserving the real subspace $\mathbb{R}^n$ of $\mathbb{C}^n$. So the fibre is the orthogonal

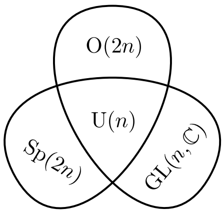


Figure 3.27: Intersection of the three subgroups in $\mathrm{GL}(2n, \mathbb{R})$.

group and hence we identify the Lagrangian Grassmannian manifold with the quotient

$$\Lambda(n) = \mathrm{U}(n)/\mathrm{O}(n)\,.$$

This fibration is helpful for the solution of our problem:

$$\dim\left(\Lambda(n)\right) = \dim\left(\mathrm{U}(n)\right) - \dim\left(\mathrm{O}(n)\right) = n^2 - \frac{n(n-1)}{2} = \frac{n(n+1)}{2}\,.$$

To calculate the fundamental group, consider the map $\det : \mathrm{U}(n) \to \mathbb{S}^1$. It is not determined over the base space of the fibration $\mathrm{U}(n) \to \Lambda(n)$, since the fibre is $\mathrm{O}(n)$ and the determinant of an orthogonal matrix may be 1 or -1.

However, the map $\det^2 : \mathrm{U}(n) \to \mathbb{S}^1$ induces an unambiguously defined map "$\det^2$" $: \Lambda(n) \to \mathbb{S}^1$, which is a smooth fibration.

Maslov index. The map $\det^2$ associates to any loop in the Lagrangian Grassmannian manifold $\Lambda(n)$ a "rotation number", which is the increment of the argument of the complex number $\det^2(g)$ along the matrices on the lifted loop in $\mathrm{U}(n)$. This integer number is called the *Maslov index* of the loop.

Theorem 17. *The Maslov index provides the homotopy classification of the loops in the Lagrangian Grassmannian manifold* $\Lambda(n)$:

$$\pi_1(\Lambda(n)) \approx \mathbb{Z}\,.$$

The details of the proof of this result are given in [5]. Arnold's lectures (Paris, 1965) on this subject were later published, adding to them new contributions, by one of the listeners, J. Leray [95] who was the advisor of Arnold during his stay in Paris.

The Maslov index has several applications and, after [5], some more general definitions of it were given (starting from Leray's one). In Subsection

16.11.5 (pp. 632-636) we give a different presentation of it and describe its natural appearance in quantum mechanics and in symplectic and contact topology of caustics and wave fronts.

3.16 Differentiable structures : Milnor spheres

The theory of smooth manifolds and the ordinary theory of manifolds (continuous, C^0-topology) are very similar in dimension 2 : every topological manifold, like a cube surface, may be smoothed in order to become a smooth surface homeomorphic to the original one, and any two non-diffeomorphic smooth surfaces are not homeomorphic. The situation in dimension 3 is the same as in dimension 2 : every topological 3-manifold has a unique smooth structure up to diffeomorphism.

However, in higher dimensions the situation is sometimes very different. There exist closed ordinary manifolds that are not homeomorphic to any smooth manifold – they are called *non-smoothable*, and can be defined by rather simple algebraic equations.

Moreover, there exist smooth manifolds which are homeomorphic between them but not diffeomorphic. They were discovered by J. Milnor. An example is provided by the 7-dimensional manifold defined in $\mathbb{C}^5$ by the equations

$$x^3 + y^5 + z^2 + u^2 + v^2 = 0 \ , \qquad |x|^2 + |y|^2 + |z|^2 + |u|^2 + |v|^2 = 1 \ .$$

This system contains three real equations because the first of the above equations means the equalities of the real and of the imaginary parts of complex numbers. Hence, they define a smooth 7-manifold in $\mathbb{C}^5$, that is, in $\mathbb{R}^{10}$.

This 7-dimensional smooth manifold is homeomorphic to the usual sphere $\mathbb{S}^7$, but is not diffeomorphic to it. The homeomorphism proof is not too difficult, it suffices to construct on the manifold a function with just two critical points, a maximum and a minimum. The proof that this manifold is not diffeomorphic to the ordinary sphere $\mathbb{S}^7$ is more difficult: One needs to define some invariants of smooth manifolds that we shall approach later.

The 7-manifold described above is called "exotic sphere" or "Milnor sphere", while the equations for it came from E. Brieskorn's studies on the degenerations of the critical points of complex functions. In fact, Milnor classified all the smooth manifolds homeomorphic to $\mathbb{S}^7$: There are 28 non-diffeomorphic such manifolds. All these manifolds can be defined by systems of equations in $\mathbb{C}^5$, similar to the above Brieskorn example.

The only difference is that one should replace the exponent 5 in Brieskorn's formula by the number $6k - 1$. Choosing $k = 1, \ldots, 28$, one gets all the 28 Milnor spheres, including a strange realisation of the usual sphere for one particular case of these 28 manifolds.

The 28 Milnor spheres form a group isomorphic to $\mathbb{Z}_{28}$, of which the standard sphere represents the zero element, in the additive notations.

The group operation for the "Milnor spheres" homeomorphic to $\mathbb{S}^n$ is defined by the following Milnor construction. Each of these manifolds can be obtained from two standard discs, like the usual sphere can be decomposed into the north hemisphere and the south hemisphere. Their gluing is defined by a map of the boundary $\mathbb{S}_+^{n-1}$ of the north part to the boundary $\mathbb{S}_-^{n-1}$ of the south part. This map should be a diffeomorphism. For the diffeomorphy between the glued manifold and the usual sphere $\mathbb{S}^n$, the gluing map should be homotopic to the identity map in the space of gluings (namely, the identical identification of both copies of the equator provides the standard sphere).

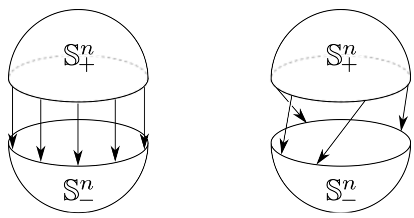


Figure 3.28: The semi-spheres gluings generating the ordinary sphere $\mathbb{S}^n$ and an exotic Milnor sphere.

The diffeomorphic type of the glued manifold remains unchanged when the gluing diffeomorphism is replaced by a close one, and hence remains the same type for all the homotopic gluings. In this way, Milnor associated his spheres to the components of the group of orientation preserving diffeomorphisms of the smaller sphere $\mathbb{S}^{n-1}$.

For the case $n = 2$ there is only one homotopy class of the circle diffeomorphisms preserving the circle orientation. Hence, there is no non-standard Milnor sphere of dimension 2.

The group structure is defined simply by the composition of the gluing diffeomorphisms: It defines a group operation between the connected components, defining hence a group operation on the resulting spheres.

Calculating the components of the gluing diffeomorphisms, the mathematicians have determined the number m_n of Milnor spheres homeomorphic to $\mathbb{S}^n$:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
m_n	1	1	1	?	1	1	28	2	8	6	992	1	3	2	16256	$\dots$.

The groups of Milnor spheres are cyclic and the largeness of $m_{11} = 992$ is explained by the arithmetics of the corresponding Bernoulli number involved in the classification of diffeomorphisms, but this interesting interrelation between so different domains of mathematics cannot be explained in the present elementary textbook.

Connected sum operation. An equivalent way to define this group operation for two spheres $\mathbb{S}^n_1$, $\mathbb{S}^n_2$ is the construction called "connected sum". One removes a small disc from each sphere and then the resulting boundaries (each diffeomorphic to the standard sphere $\mathbb{S}^{n-1}$) are glued by the standard diffeomorphism (taking into account the orientation). The manifold we obtain is the *connected sum* $\mathbb{S}^n_1 \# \mathbb{S}^n_2$.

On the differentiable structures of $\mathbb{R}^4$. The number of different non diffeomorphic smooth manifolds that are homeomorphic to the standard Euclidean 4-space $\mathbb{R}^4$ is infinite.

Every such manifold can be obtained from the standard Euclidean space $\mathbb{R}^5$, equipped with a suitable smooth vector field v with no zeros.

Namely, the space of orbits of such a field should be a smooth manifold homeomorphic to $\mathbb{R}^4$. But these manifolds associated to different vector fields are generically non diffeomorphic to each other.

It is unknown, however, how to write explicitly a vector field v for which the orbit manifold is non diffeomorphic to the standard space $\mathbb{R}^4$.

Question. Such a vector field is defined by five functions. May them be polynomial, trigonometric or analytic ? How to write them ?

3.16.1 Smooth, topological and polyhedral manifolds

Concerning the problem of the classification of manifolds, we notice that one ought to distinguish even three different topological theories of manifolds: The smooth topology Diff, the ordinary topology Top and the polyhedral topology PL (for "piece-wise linear").

The last theory consider polyhedral manifolds, and the equivalence homeomorphisms should be linear on its faces.

The legend says that Poincaré invented topology for the following reason. The examiners of the École Normale Supérieure in Paris had underestimated his examination paper in which the circles of the figures were too similar to the triangles, to which they are only PL-equivalent, but not diffeomorphic.

To fight against this discrimination, Poincaré finished the Parisian École Polytechnique rather than the École Normale Supérieure (some kind of MIT, replacing Harvard) and — according to the legend— invented topology.

The examination part of this legend is certainly true. Poincaré was graduated in the Parisian École de Mines where one can see his written accounts of the Swedish mining, done together with another student, who produced the drawings.

Later his friends advised him to start celestial mechanics works, since they expected an academy member place to be free in the astronomical department: The idea was that in the mathematics department his teacher Hermite liked too much to correct triangles and circles.

In dimension 2 all the three theories provide the same classification of surfaces, but all the three higher-dimensional theories are in general different. The PL-theory seems to be intermediate between Top and Diff, but it provides in fact a classification different from both others.

These three categories of objects had been mentioned already by Lucretius ("De rerum natura") two millenniums ago. Namely, he proposed to explain the different chemical properties of different atoms by the peculiarities of their surfaces. He suggested just three different types: Smooth surfaces (providing good sliding of one atom along the other, reducing to weak chemical activity for the generation of molecules), polyhedral surfaces (providing worse sliding, due to the angles obstructing it) and coarse surfaces (where small hooks can connect atoms, completely obstructing any sliding and providing strong chemical relations).

3.17 Topology simplification in higher dimensions

As we have seen, the topology can change drastically with the dimension of the objects. We have to observe that a general lesson of the topology development in the 20th Century is that most situations become simpler in higher dimensions, and that the infinite dimension limit could be the simplest thing to study.

Example. The classification of knots (that is, of embeddings of the circle in Euclidean 3-space, considered up to homotopy of the embeddings) is a very difficult part of topology and the general classification of knots is still unknown in spite of the attempts to study it. Such attempts were started by Kelvin, who expected to obtain some classification of the different atoms by classifying the knots (we discuss about knot invariants in Chapter 18).

In dimension four the knot "disappear": Every embedding $\mathbb{S}^1 \to \mathbb{R}^4$ is homotopic to any other. For instance, to the standard embedding of the circle in a plane of $\mathbb{R}^4$.

This topological fact is true in the three categories Diff, Top and PL, and is not too difficult to prove it as an exercise. The spiritualists of the 19th Century were aware of this topological theorem on the 4-dimensional unknottedness of all knots, and claimed that they were practically using the 4th dimension in their (in fact, cheating) experiments.

Example. As we have seen above, the fundamental group of the orthogonal group of $\mathbb{R}^n$ contains two elements for $n > 2$, $\pi_1(\mathrm{O}(n)) = \mathbb{Z}_2$, while for $n = 2$ this fundamental group is infinite, confirming the topology simplification in higher dimensions.

Problem. Show that if a closed smooth m-dimensional manifold M is the boundary of a smooth $(m+1)$-dimensional manifold N, then any embedding of M in Euclidean space $\mathbb{R}^n$ can prolonged to an embedding of N, provided that n is sufficiently large with respect to m.
What means "sufficiently large" here ? (Hint: see Gibbs principle, p. 21.)

Example. When the submanifold M is a curve, $m = 1$, the most difficult case is $n = 3$, when the curve is generically knotted, being not the boundary of an embedded disc. It is however always the boundary of some orientable smooth compact surface, called a *Seifert surface* of that knot. For the trefoil knot this surface is already rather complicated (see Fig. 3.29).

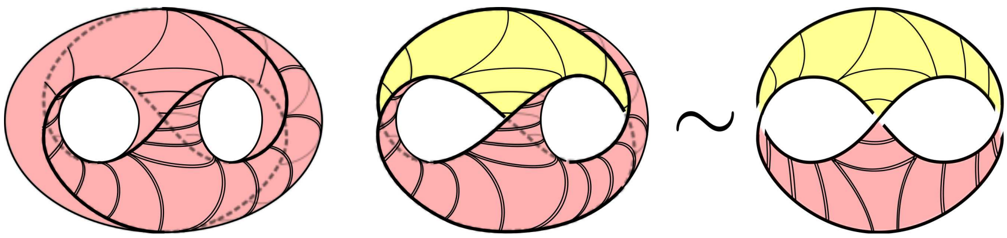


Figure 3.29: Seifert two-sided surface of the trefoil knot ($g = 2$).

The proof of the general result formulated above, for a sufficiently large value of n, is much simpler than the proof of the Seifert surface existence, confirming once more the simplification of topology in higher dimensions.

A single knot can have many different nonequivalent Seifert surfaces – adding a handle to a Seifert surface we get a new Seifert surface.

The *genus of a knot* is defined as the minimal genus of all Seifert surfaces for that knot. So the genus is a knot invariant.

In the left part of Fig. 3.29, the simplest Seifert surface of the trefoil knot is embedded as a part of the embedded closed surface of genus $g = 2$, a sphere with two handles. However, the genus of this Seifert surface, and hence of the trefoil knot is $g = 1$. Since this Seifert surface is diffeomorphic to the complement of a disc in a torus, its Euler characteristic is equal to -1.

Remark. This abstract torus, obtainable from the Seifert surface by attaching a disc along the border line, cannot be an embedded continuation of the Seifert surface, since otherwise the trefoil knot would be contractible along the attached embedded disc.

For this reason the Seifert surface of genus 1 is represented in Fig. 3.29 as a domain of an embedded closed surface of genus 2.

Remark. The trefoil knot is also the boundary of a non orientable embedded smooth surface. Namely, such non orientable surface is a special – very simple – embedding of the Möbius band.

Exercise. Draw a picture of the special embedding of the Möbius band for which the boundary is the trefoil knot.

Chapter 4

Theorem of Weierstrass and convergence theories

4.1 The Weierstrass Theorem

We have studied the fundamental group $\pi_1(M, *)$, which counts the closed curves in M starting at a chosen point $*$ (up to homotopies of the maps $\varphi : I \to M$ sending the endpoints 0 and 1 of the segments $\{0 \leqslant t \leqslant 1\}$ to the starting point $*$: $\varphi(0) = \varphi(1) = *$ – Fig. 4.1).

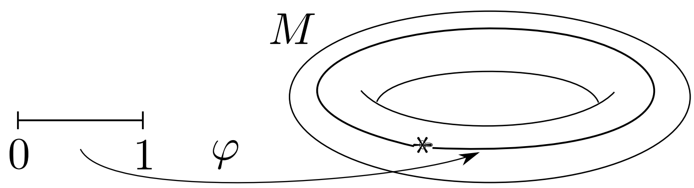


Figure 4.1: An element of the fundamental group of M, represented by a loop φ.

We have seen that $\pi_1(\mathbb{S}^1, *) = \mathbb{Z}$, the homotopy class describing the number of the turns around the (oriented) circle $\mathbb{S}^1$ (for the map $t \mapsto e^{2\pi ikt}$ the number of turns is $k \in \mathbb{Z}$ – Fig. 4.2).

For the higher-dimensional spheres $\mathbb{S}^n$ $(n > 1)$, the fundamental group is trivial. Indeed, any continuous map φ may be approximated by a smooth map $\widetilde{\varphi}$ that is homotopic to φ (for a map to a vector space, the homotopy is provided by a linear interpolation $F(s,t) = \psi_t(s) = \varphi(t) + s(\widetilde{\varphi}(t) - \varphi(t))$ – Fig. 4.3).

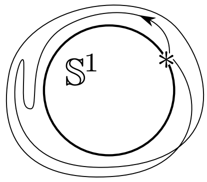


Figure 4.2: An element of the fundamental group of a circle, corresponding to the integer 2.

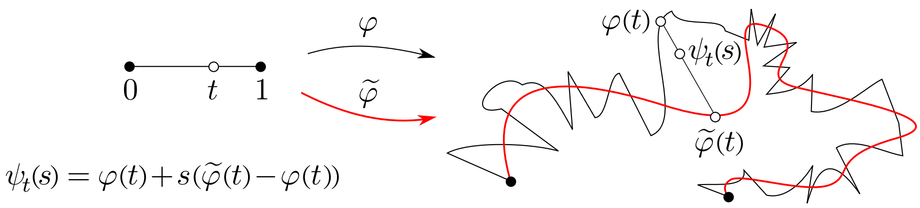


Figure 4.3: The homotopy of a continuous map φ to a smooth map $\widetilde{\varphi}$.

The image $\widetilde{\varphi}(I)$ of a smooth map from the segment I to $\mathbb{S}^n$ does not cover all the sphere (since this image has measure 0 by the Sard lemma). By an explicit homotopy of the sphere we may reduce the map $\widetilde{\varphi}$ to a homotopic map $\widehat{\varphi}$, whose image does not cover the point N, opposite to the basic point $*$ (we shall call $*$ the South pole and N the North pole – Fig. 4.4).

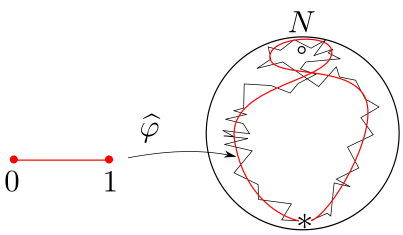


Figure 4.4: Homotopy of a loop φ to a pole-avoiding loop $\widehat{\varphi}$.

Now it is easy to make a homotopy of the sphere $\mathbb{S}^n$ to itself, preserving the South pole $*$ and the North pole N, starting from the identity map, moving each point along the meridian to the South, and ending (at time 1) by a map contracting all the complement to a small neighbourhood of the North pole to the South pole $*$ – Fig. 4.5.

The whole curve $\widehat{\varphi}(I)$ being sent to the South pole, we have constructed a homotopy $\varphi \sim \widetilde{\varphi} \sim \widehat{\varphi} \sim \psi$ of the initial map φ to the trivial map ψ (sending

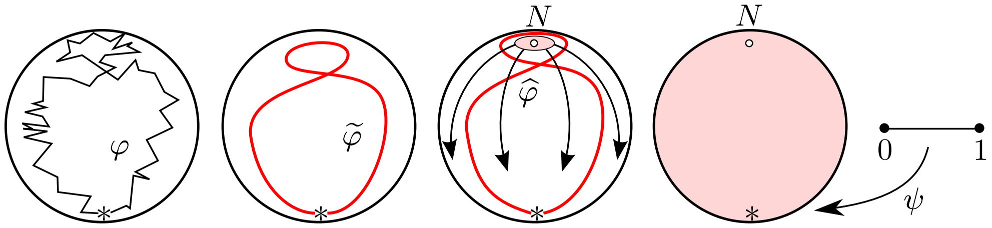


Figure 4.5: Homotopy of a loop φ to the trivial loop ψ.

all the segment I to the base point $*$).

We have thus calculated the homotopy group $\pi_1(\mathbb{S}^n, *) = 1$, $n > 1$ (we denote the trivial group, consisting of one element, by the multiplicative notation 1, rather than by the additive notation 0, since the fundamental group is, in general, non-commutative).

Example. Consider the complement $M = \mathbb{R}^2 \setminus \mathbb{S}^0$ to 2 points of the plane – Fig. 4.6.

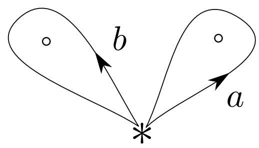


Figure 4.6: Generators a and b of the fundamental group of $\mathbb{R}^2 \setminus \mathbb{S}^0$.

The fundamental group of this surface is generated by the two loops a and b, shown in Fig. 4.6. The general element is homotopic to the product of these generators and of their inverted versions a^{-1}, b^{-1}:

$$\varphi \sim a^{\alpha_1} b^{\beta_1} a^{\alpha_2} b^{\beta_2} \dots a^{\alpha_n} b^{\beta_n} ,$$

where $\alpha_k \in \mathbb{Z} \setminus 0$ for $1 < k \leqslant n$, $\beta_k \in \mathbb{Z} \setminus 0$ for $1 \leqslant k < n$ (α_1 and β_n may also be 0, since the initial loop may be of b type, and the final of a type).

There are $4 \cdot 3^{m-1}$ words of length m, which are all different elements of the fundamental group $\pi_1(M, *)$.

This group, called *the free group* with generators (a, b), is very large. Moreover, this exponential growth of the number of elements of length m in a free group makes difficult to recognise the identities between elements in the general non-commutative groups defined by their generators and some basic relations.

To prove the above statement, that $\pi_1(\mathbb{R}^2 \setminus \mathbb{S}^0, *)$ is free, we first represent the loop $\varphi : I \to (\mathbb{R}^2 \setminus \mathbb{S}^0)$ by a smooth loop $\widetilde{\varphi}$, homotopic to it, making it also transverse to the straight paths A and B, connecting the deleted points of $\mathbb{S}^0$ to infinity – Fig. 4.7.

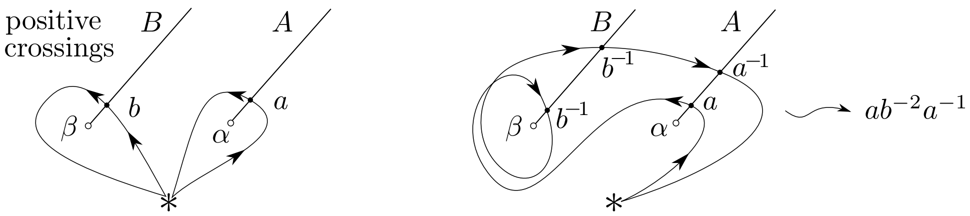


Figure 4.7: Construction of the word from a given loop.

We write a or a^{-1} for the crossing of A (in the positive or in the negative direction), b or b^{-1} for the crossing of B, and we obtain the word homotopic to φ.

We have thus proved that $\pi_1(M, *)$ is generated by the 2 elements a and b. To prove that it is free (that is, that different words represent different elements of the fundamental group) one might try to do it as a problem. We shall provide a formal proof later, using covering theory.

Remark. We have used in the proofs above the Weierstrass approximation theory, deforming a continuous function on I to a smooth one.

The proofs of this (difficult) theorem are in general contained in the calculus textbooks. See, for instance, a nice exposition of this theorem in the book "A Mathematician's Miscellany" of J. Littlewood, starting from an unreadable version by a student and explaining how an experimented teacher should make the same proof understandable.*

However, the story of this discovery is usually hidden from the students (to enhance professors' authority), and we shall explain it now.

---

*We add an astonishing extension of the Weierstrass theorem (which we shall not use). The polynomials of one variable x are the linear combinations of the monomials x^a $(a = 0, 1, 2, \ldots)$.

Question. Which sets of the monomials of degrees 0 and a_k approximate all the continuous functions (on a segment) by their linear combinations?

Answer. *For this C^0-density of the linear combinations of the monomials of degrees a_k it is necessary and sufficient the divergence of the series $\sum(1/a_k) = \infty$.*

This result has been first discovered by S.N. Bernstein and this conjecture of Bernstein was then proved by H. Müntz, 1912.

Newton, using the Archimedes limits, never formulated the exact definitions (of $\varepsilon \to \delta$ or $\varepsilon \to n$ type), that we all know today. Cauchy was the first to introduce them in the formal teaching, and one of the first theorems of his "rigorous" calculus course stated that:

If the functions $f_n : I \to \mathbb{R}$ are continuous, the limiting function f, defined by the limits

$$f(x) = \lim_{n\to\infty} f_n(x) \ ,$$

is continuous too (provided that the above limits do exist for every $x \in I$).

While Cauchy published his "rigorous" ε-proof of this basic theorem, the young Norwegian* student Abel provided easy counterexamples (Fig. 4.8), like even $f_n(x) = x^n$ (where $\lim_{n\to\infty} x^n = 0$ for any $0 \leqslant x < 1$, but $\lim_{n\to\infty} x^n = 1$ for $x = 1$, making f discontinuous).

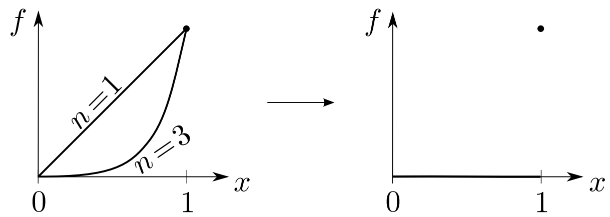


Figure 4.8: Point-wise convergence of the continuous functions f_n to a discontinuous function f.

He showed his counter-example to Cauchy, but according to Abel's letters "French mathematicians were all inclined to teach their science rather than to study anything new".

The result was that the Paris Academy of Sciences lost the manuscript of Abel where he had proved the impossibility of solving algebraic equations of degree 5 or higher by radicals (the topological content of this algebraic theorem is explained in chapter 13).

Abel's manuscript had been sent by the Academy (for a scientific review) to Cauchy, who never answered. Later he acted similarly in the case of the Galois paper on the impossibility of the similar arithmetical expressions (for some particular values of the parameters).

*Abel had spent a year in Paris and after his return to Christiania the Paris newspaper claimed that this young man was so poor that he was obliged to return to "his part of Siberia, called Norway" by foot on the Atlantic ice.

Fortunately, Abel's manuscript (as well as that of Galois) was rediscovered in the Cauchy papers after the death of this great mathematician. Liouville published them, and these theories of Abel and of Galois created the most flourishing parts of mathematics of the XIX Century (group theory, Riemannian surfaces theory). However, it happened many decades later, while both Abel and Galois died young, being unappreciated by their contemporaneous.

It is astonishing how stable are these characteristics of the scientific community – be it the world mathematics, or some particular countries traditions.

Anyway, in natural sciences and geometry most questions on Cauchy's "point-wise limits" f should be be replaced by the "uniform convergence" (where the decline $|f_n(x) - f(x)|$ should be smaller than the same ε simultaneously for all the points $x \in I$).

In the case of the uniform convergence, the limiting function of a convergent sequence of continuous functions is indeed continuous itself. The really rigorous proof of this "Weierstrass theorem" is described in all the modern textbooks (replacing the Cauchy wrong theorem and wrong proof, which should in fact provide to the students more understanding of the limit theory than the ε-δ definition).

It is interesting that, being uniformly approximable by smooth maps or by polynomials, continuous functions and maps behave sometimes very differently from the naive expectations.

Thus, there exist continuous functions that are *nowhere* differentiable. Examples can be constructed by hand by means of approximating broken lines (Fig. 4.9) with segments having fastly varying slopes. Weierstrass wrote the following explicit formula for such a construction: $f(x) = \sum_0^\infty a^n \cos(b^n \pi x)$, where $0 < a < 1$ and b being any positive integer satisfying $ab > 1 + 3\pi/2$.

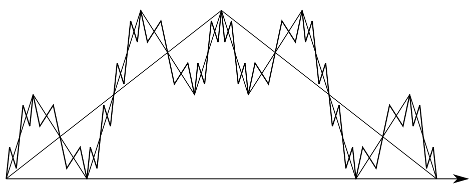


Figure 4.9: Nowhere differentiable continuous function.

For a continuous map of the segment I to the plane, the image may cover a whole domain (like, say, the square I^2). To construct such a strange map (called a "Peano curve"), it suffices to consider the uniform limit of the broken lines shown in Fig. 4.10: when n is growing, the approximating broken line is visiting all the points of I^2 ε-neighbourhoods, and the Weierstrass uniform limit of this sequence is a continuous map $\varphi : I \to I^2$, covering the whole square ($\varphi(I) = I^2$).

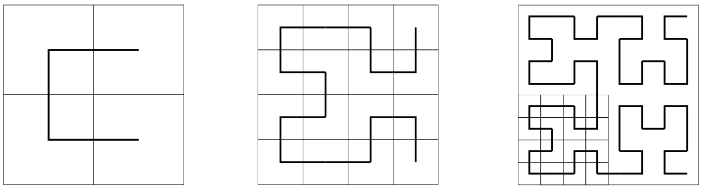


Figure 4.10: Peano curve approximations.

Similarly, there exist maps covering the cube I^m of any dimension m.

Definition. A continuous map f verifies *the Hölder condition* α, if there exist $C > 0$ such that for any $\varepsilon > 0$

$$|f(x_1) - f(x_2)| < C\varepsilon^{\alpha}$$

for all the pairs of arguments for which $|x_1 - x_2| < \varepsilon$.

In the case $\alpha = 1$ the Hölder condition is called Lipshitz condition (C being the estimation of the maximum of the modulus of first derivative of f on I). In the case $\alpha > 1$ the Hölder condition is verified only by the constant functions.

Problem. Construct a Peano map $f : I \to I^m$, covering the cube I^m, verifying the Hölder condition with $\alpha = 1/m$, and prove that such a map can not be "smoother": if $f : I \to I^m$ verifies the Hölder condition with $\alpha > 1/m$, then the image can not cover the whole cube: $f(I) \neq I^m$.

Similarly, $\alpha \leqslant \frac{n}{m}$ for a map $f : I^n \to I^m$, covering I^m.

Before we leave this subject of convergence theories, we shall mention two more phenomena of natural sciences, usually omitted in the calculus textbooks, but extremely important for many scientific and geometric applications of calculus.

Strangely, these facts were discovered and studied rather by some experts in applications than by mathematicians. While they belong to the very core of the basic mathematical convergence theory, they are mostly hidden from the students of mathematics (perhaps, to enhance the mathematics authority in their naive souls).

4.2 The Gibbs phenomenon and tomography

The first phenomenon is called *the Gibbs phenomenon*[*] – Fig. 4.11.

The observation is that *the graph of the limiting function can be different from the limit of the graphs of the functions of the series* (which converge to the limiting function at every point).

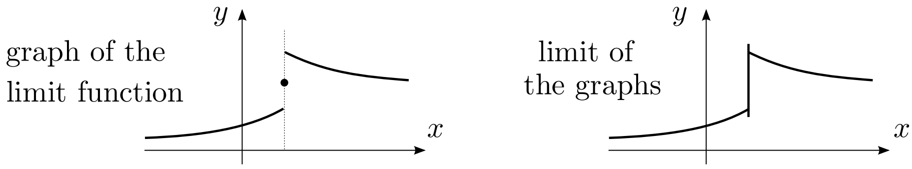


Figure 4.11: Gibbs phenomenon.

Gibbs observed this experimentally, studying the convergence of Fourier series, for practical works on tides prediction.

If the function, for which we study the Fourier series, has a discontinuity, the series still converges (provided that it is piecewise smooth in the complement to the discontinuity point). The classical Fourier series theory, exposed in the textbooks, claims that the series limit is equal to the initial function in every interval that does not contain a discontinuity, while at a discontinuity point the value of the limiting function equals the arithmetical mean of the left and of the right branches of the function – Fig. 4.11 left.

The Gibbs observation was that, in spite of the correctness of this deep theorem of Fourier series theory, the practical measurements in this nonuniform convergence case are quite different. The limit of the graphs of the sums of finite numbers of terms of the series contains (in addition to the graph of the limit-function) a vertical segment connecting the two discontinuity branches. This connecting segment is prolonged (above and below the connected branches) by two standard appendices whose lengths are (always) approximately 9% longer than the connecting segment – Fig. 4.12.

Unfortunately, this phenomenon is practically never described in the Fourier series expositions in the calculus textbooks. The only exception that we know is the R. Courant textbook.

The discovery of the Gibbs phenomenon description in Courant's book has made on V. Arnold (who was then, perhaps, 15 years old) a very deep

[*]Gibbs was the first great American mathematician, known more for his works on the applications of geometry to thermodynamics and to the bases of the statistical physics.

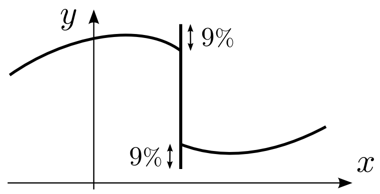


Figure 4.12: Gibbs appendices.

impression, showing him the important difference between the mathematical theories and the real world phenomena that they wish represent.

Next Arnold's collision with the Gibbs phenomenon came from a very different subject – the study of the tomographical images:

Radon Transform The interior density $f(x,y)$ of the body of the patient along a chosen plane $z =$ const is studied in terms of the "absorption" of the rays crossing the body in this plane. Each ray is characterised by its direction φ and its distance q from the "central ray" (Fig. 4.13), and the

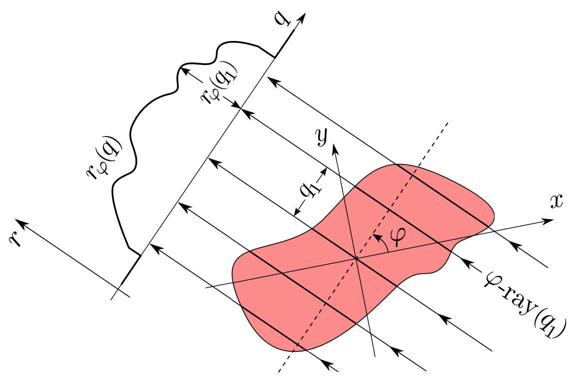


Figure 4.13: Radon transform in tomographical measurements.

resulting absorption provides the description of the integral

$$r(\varphi, q) = \int_{\varphi\text{-ray}(q)} f\big(x(t), y(t)\big)\; dt$$

along the ray $\{x(t), y(t)\}$ corresponding to the parameter values φ and q.

This function r is called the *Radon transform* of the studied density f. The problem is to calculate the function f of two variables (x,y) from the measured values of the Radon transform function r (which also depends on two arguments, q and φ).

This Radon transform inversion problem is solved by the computer, using Fourier theory. Indeed, the Fourier transform of the function r in the variable q

$$\widehat{r}(\varphi, p) = \int r(\varphi, q)\ e^{i(p,q)}\ dq$$

is just the Fourier transform of the initial function f (at the corresponding wave vector, defined by the direction φ and the position p).

Thus, calculating the "sum" (instead of the integral that we have written) of the Fourier series representation of f (whose "Fourier coefficients" are $\widehat{r}$), the computer provides the unknown interior density f.

Doing all this, the medical experts were astonished to find, together with the known changes of f at the boundaries of different interior tissues of the patients, some additional features, corresponding to no biological structure.

Gibbs rings artefacts. Such objects as the bones are well seen, their boundaries representing some discontinuities of the density function f (along some curved lines in the plane with coordinates x and y). Near these curves, the computer adds some "ghosts", called *Gibbs' rings*, which, visually, appear as alternating bright and dark blurring bands parallel and adjacent to the borders of an area of abrupt density change (or signal intensity change), such as the borders between cortical bone and surrounding fat (like in Fig. 4.14).

Figure 4.14: Gibbs ringing artefacts.

The only way to explain these strange artefacts is to take into account the Gibbs phenomenon at the discontinuity represented by the bone boundary.

The "Fourier series summation" provides the Gibbs appendices in the discontinuities of each density function $r_\varphi(q)$, at the tangent rays, which, in our two-dimensional case, could take place on the whole discontinuity curve.

Gibbs rings increase by insufficient collection of samples and simply because the Fourier series (used to reconstruct the signal) are "truncated".

Tangential artefacts. The computer can also add to the bone boundary some straight lines (where nothing happens in the body) – the double tangent lines to the bones boundary curve and the inflectional tangents – Fig. 4.15.

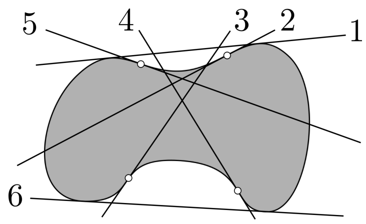


Figure 4.15: Tomographical artefacts 1-6.

These tangential artifacts are due to the singularities of the dual curve of the bone boundary (Fig. **??**) combined with the beams' discreteness. In fact, the Radon transform is a function defined in the two-dimensional space formed by all the lines of the plane (the rays). In this space (of lines), the points of the dual curve of the bone boundary are the tangent lines to this boundary. Singularities of dual curves (and hypersurfaces) belong to projective duality and, hence, to contact geometry (see p.**??**).

Thus, the understanding of Gibbs phenomenon and dual singularities are quite necessary for dealing with tomography artefacts (which could otherwise lead to disastrous medical intervention of ignorant doctors).

Similar phenomena should occur in seismology, where the absorption of the seismic waves (produced, say, by earthquakes, by volcanic activity or by artificial explosions) by the interior body of the planet is used to calculate the interior density, knowing the observations at different points of the planet surface at different moments of time.

The computer provides (by tomography type calculations) some interior densities $f(x, y, z, t)$, whose discontinuities (say, on the geophysical boundaries) are extremely important (both for the theoretical study of the interior structure and for the oil research perspectives).

The Gibbs phenomenon should provide here some "artefacts", similar to the medical bones tangent lines, which the geophysicists should distinguish from the real discontinuity of the physical world.

4.3 Weak convergence

Another – very different – type of convergence is usually called "weak convergence", and it describes better the geometrical and physical reality than the point-wise convergence and the uniform convergence in many practical problems.

It is related to turbulences and fast oscillations of many objects (both in space and in time). The problem is to study averages rather than fluctuations, and the mathematical technology of it is not so easy in many cases.

The simplest example is the sequence of the fastly oscillating harmonics, $f_k(x) = \cos(kx)$ – Fig. 4.16.

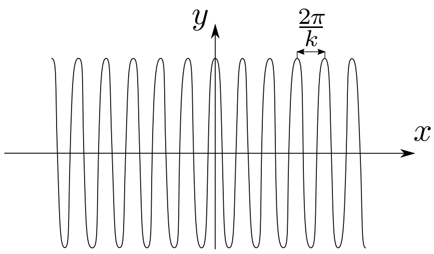


Figure 4.16: Weakly small function.

While the values of the cosines are somewhere $+1$ and somewhere -1, these contributions "annihilate" each other, and the "practical" effect is close to zero:

$$\int g(x) f_k(x)\ dx \to 0$$

for $k \to \infty$ and for any smooth function g (which we wish to evaluate, measuring its influence by f_k).

In such cases one says that the sequence f_k is *weakly convergent to zero* for $k \to \infty$. For the sums $F + f_k$ (with a fixed, say smooth function F) one wishes to consider their weak convergence to F for $k \to \infty$. Given some practically measured quantity, in which the "regular part" F and the "oscillating part" f_k are not separated a priori, it is sometimes difficult to calculate the weak limit.

Knowing a lot of examples where it had been done successfully, we quote in the next section one different case where the problem remains still open.

4.4 Distributions of the Frobenius numbers

Let $a_1, \dots, a_n$ be positive integers having no common divisor greater than 1. It is not too difficult to prove that any sufficiently large integer N can be represented as a linear combination of them

$$N = p_1 a_1 + \cdots + p_n a_n \,,$$

where the "multiplicities" p_m are also non-negative integers.

Sylvester proved that for two coprime integers ($n = 2$) this representation is possible for all $N \geqslant K(a) = (a_1 - 1)(a_2 - 1)$ and the product K is the minimal integer having this property. To understand it, consider the

Example. For $n = 2$ with $a_1 = 3$, $a_2 = 5$, we have $K = (3-1)(5-1) = 8$. In this case it is easy to represent the numbers

$$N = 3, 5, 6, 8, 9, 10 \,.$$

All numbers $N > 10$ are represented because adding 3 to 8, to 9 and to 10 we get the three consecutive numbers 11, 12, 13, and one can always add 3.

Sylvester's Remark. Up to the Sylvester value K one half of the numbers are representable and the other half is not. The distribution of both halves is symmetrical: for $x + y = K - 1$, x is representable if and only if y is not.

The problem to evaluate the minimal number $K(a_1, \dots, a_n)$ is not solved even for three initial numbers ($n = 3$), and the experimental results oscillate wildly with the vector $\boldsymbol{a}$. Consider the following partitions $a + b + c = 23$:

The trapezium of partitions $a + b + c = 23$,

$(a, b, c) =$

9,2,12	9,3,11	9,4,10	9,5,9	9,6,8	9,7,7			
8,2,13	8,3,12	8,4,11	8,5,10	8,6,9	8,7,8			
7,2,14	7,3,13	7,4,12	7,5,11	7,6,10	7,7,9	7,8,8		
6,2,15	6,3,14	6,4,13	6,5,12	6,6,11	6,7,10	6,8,9		
5,2,16	5,3,15	5,4,14	5,5,13	5,6,12	5,7,11	5,8,10	5,9,9	
4,2,17	4,3,16	4,4,15	4,5,14	4,6,13	4,7,12	4,8,11	4,9,10	
3,2,18	3,3,17	3,4,16	3,5,15	3,6,14	3,7,13	3,8,12	3,9,11	3,10,10
2,2,19	2,3,18	2,4,17	2,5,16	2,6,15	2,7,14	2,8,13	2,9,12	2,10,11

The values of $K(a, b, c)$ are represented in Fig. 4.17 for these partitions.

The conjectural "weak limit" answer for $\boldsymbol{a} \to \infty$ is the growing rate

$$K \sim K_0 = (C(a_1 \ldots a_n))^{1/(n-1)} \,,$$

with the coefficient $C = (n-1)!$. In the case $n = 2$ it provides the weak asymptotics

$$K \sim a_1 a_2 \ ,$$

confirmed by the Sylvester formula.

$K(4,6,13) = 16$,

$K_0(4,6,13) = \sqrt{624} \approx 25$;

$$K(a,b,c) = \begin{array}{c}
8 \quad 20 \quad 16 \quad 32 \quad 20 \quad 48 \quad 20 \\
12 \quad 14 \quad 30 \quad 28 \quad 20 \quad 42 \quad 42 \\
6 \quad 12 \quad 18 \quad 14 \quad 16 \quad 48 \quad 42 \quad 48 \\
14 \quad 26 \quad 16 \quad 20 \quad 50 \quad 16 \quad 20 \quad 20 \\
4 \quad 8 \quad 12 \quad 48 \quad 20 \quad 14 \quad 28 \quad 32 \quad 28 \\
16 \quad 6 \quad 42 \quad 12 \quad 16 \quad 18 \quad 30 \quad 16 \quad 16 \\
2 \quad 32 \quad 6 \quad 8 \quad 26 \quad 12 \quad 14 \quad 20 \quad 18 \quad 20 \\
18 \quad 2 \quad 16 \quad 4 \quad 14 \quad 6 \quad 12 \quad 8 \quad 10 \quad 10
\end{array}$$

the radius $r = 1$ mean value:

$$\frac{K(4,6,13) + K(4,5,14) + K(5,5,13) + K(5,6,12) + K(4,7,12) + K(3,7,18) + K(3,6,14)}{7} = 22.$$

Figure 4.17: Table of Frobenius problem answers $K(a,b,c)$ for $a+b+c = 23$.

In the case $n = 3$ the proposed asymptotics is

$$K \sim \left(K_0 = \sqrt{2abc}\right) \ , \tag{1}$$

and it should be compared with the observed values of Fig. 4.17.

The thing to confirm is the approximation of the averages (along, say, the neighbourhoods of radius r of the vector $\boldsymbol{a}$) of the observed quantities $K(\boldsymbol{a})$ to the proposed quantities (1), in the sense of the relative error ($K/K_0 \to 1$ for $\boldsymbol{a} \to \infty$).

To write the general definition of the weak asymptotics, one has to average along the neighbourhood of radius $r(|\boldsymbol{a}|)$ of the vector $\boldsymbol{a}$, where $r \to \infty$ but $r/|\boldsymbol{a}| \to 0$ for $|\boldsymbol{a}| \to \infty$: the choice of $r = \sqrt{|\boldsymbol{a}|}$ suffices in many cases.

The same geometric reasoning for the weak asymptotic of the physical averaging, which provides the suggestion (1), leads to the following extension of the above Sylvester remark that one half of the numbers N between 0 and K are representable ("belong to the additive semigroup $\{p_1 a_1 + \cdots + p_n a_n\}$") in the case of two initial numbers, $n = 2$.

Namely, the expected quantity of the semigroup members in the interval $[0, K_0]$ is expected to be in many cases (weakly!) $1/n$ of K_0. The geometry provides even more: the "density" of the semigroup of representable numbers N in the interval should be (weakly) of order – Fig. 4.18

$$t^{n-1}\, dt \qquad (\text{for } t = N/K_0) \ ,$$

and the part $1/n$ follows from the integral value (see more details in [32])

$$\int_0^1 t^{n-1}\, dt = \frac{1}{n}\ .$$

The main idea is to study the simplex S in $\mathbb{R}^n$:

$$S(\ell) = \{(p_1 \geqslant 0, \dots, p_n \geqslant 0) : a_1 p_1 + \dots + a_n p_n \leqslant \ell\}.$$

Its volume equals $V(\ell) = \dfrac{1}{n!}\Pi(\ell/a_k) = \dfrac{\ell^n}{n!}\Pi(1/a_k)$. The number $I(\ell)$ of integral points p inside $S(\ell)$ grows like $V(\ell)$, whence the covering condition $dV/d\ell \geqslant 1$, for $\ell \geqslant \ell_0$, which provides $\ell_0 = K_0 = ((n-1)!\,\Pi a_k)^{1/(n-1)}$.

density of the number of representations of N

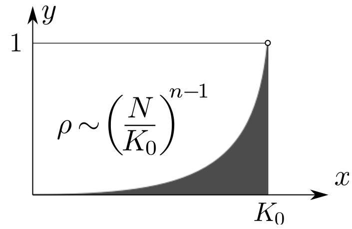


Figure 4.18: The semigroup density distribution.

There are some exceptions for those statements on the weak limits, and the number theoretical conjecture that the density of these exceptional arguments in the space of vectors $\boldsymbol{a}$ is small when $\boldsymbol{a} \to \infty$ is not proved.

To formulate this conjecture, we start by saying that an integer between 0 and $N-1$ (including both ends) is of (+)-type, if it belongs to the additive semigroup of the sums $p_1 a_1 + \dots p_n a_n$ with non-negative integer coefficients p_i, and otherwise it is of $(-)$-type. Thus, 0 is of (+)-type, while $N-1$ is of $(-)$-type.

If x is of (+)-type, then its dual integer $y = N-1-x$ is of $(-)$-type, since otherwise $N-1$ would belong to the semigroup. However, sometimes both dual numbers (x and $y = N-1-x$) are of $(-)$-type and, in this case, such integers are said to be of $(-,-)$-type.

Write x_- for the minimal value of x that belongs to the $(-,-)$-type and x_+ the maximal in the $(-,-)$-type. The conjecture says that $x_+ - x_-$ belongs to the semigroup, and that the other values of x belonging to the $(-,-)$-type have the form $x_- + q_1 a_1 + \dots + q_n a_n$, where the coefficient q_i is non negative and is less than or equal to the coefficient Q_i determined by the value $x_+ = x_- + Q_1 a_1 + \ldots + Q_n a_n$.

This "parallelepiped" of $(-,-)$-type points cannot be too large, since no large parallelepiped is contained in a simplex of a given volume: The maximal volume of a parallelepiped in the simplex of volume one in $\mathbb{R}^n$ is $n!/n^n$, which in $\mathbb{R}^3$ gives $2/9$.

Thus on gets some upper bound for the number of integers of $(-,-)$-type, and hence some lower bound for the number of integers of $(+)$-type.

Anyway, the existence of any pair (x, y) of dual numbers of $(-,-)$-type is statistically an exceptional case: Conjecturally, such pairs are absent for most part of (large) vectors $\boldsymbol{a}$.

4.5 Distributions

The physical convergence is better described in mathematical terms as the convergence of distributions, rather than as limits of functions, specially when the convergent objects are not smooth. A typical example is the "δ-function" of Dirac, defined as a non-negative function whose value is 0 at each point $x \in \mathbb{R}$ and whose integral along the real line is equal to 1.

There is no such function in mathematics, but it can be considered as some "generalised function" (or "distribution") to which a sequence of ordinary smooth functions converges, for some special understanding of the convergence that is different both from the point-wise convergence of Cauchy and from the uniform convergence of Weierstrass, but better describing the physical convergence studied by Newton, Stokes, Poincaré and Dirac.

Chapter 5

Geometry of fundamental groups

One nice example of a large fundamental group is provided by the so-called braid groups.

5.1 Braid groups

We start from an unordered set of n points of the usual plane $\mathbb{R}^2$. All the configurations of n different points of the plane form a smooth manifold of dimension $2n$. One may describe it using the direct product of n copies of the plane, $\mathbb{R}^{2n} = (\mathbb{R}^2)^n$, considering it as the configuration space of n points which are ordered and can coincide. To avoid the coincidence we eliminate the diagonals (the codimension 2 subspaces $\Delta_{k,m} \approx \mathbb{R}^{2n-2}$, defined by the equations of the form $x_k = x_m$ for the ordered points $x_k \in \mathbb{R}^2_k$, $x_m \in \mathbb{R}^2_m$).

The open domain $\mathbb{R}^{2n} \setminus \cup_{k<m}\Delta_{k,m}$ is the configuration manifold of n ordered different points of the plane. The permutation group $S(n)$ acts on it, permuting the points arbitrarily. The quotient manifold

$$\mathrm{Conf}_n(\mathbb{R}^2) = \frac{\mathbb{R}^{2n} \setminus \bigcup_{k<m} \Delta_{k,m}}{S(n)} ,$$

called nth *configuration space* of the plane $\mathbb{R}^2$, is still a smooth manifold.

Braid Group. The *group of braids with n-strings*, $\mathrm{Br}(n)$, is the fundamental group of the n-th configuration space of the plane,

$$\mathrm{Br}(n) = \pi_1 \left(\mathrm{Conf}_n(\mathbb{R}^2), *\right) .$$

The name “braid” comes from the following interpretation (Fig. 5.1).

The braid associated to a loop $\varphi : I \to \mathrm{Conf}_n(\mathbb{R}^2)$ consists of n curves in $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R}^2$, which are n continuous graphs, called *strings*:

$$(t, x_1(t))\,,\ (t, x_2(t))\,,\ \ldots\,,\ (t, x_n(t))\ .$$

The points of the configuration are ordered at the initial moment $t = 0$ and the branches x_k are defined for $0 < t \leqslant 1$ by continuity; different strings never intersect and, being graphs, the strings follow the t-direction of $\mathbb{R}^3$ positively.

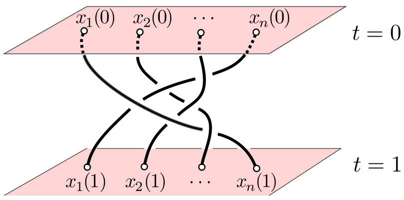


Figure 5.1: The n strings of a braid.

The ending configuration $\{x_k(1)\}$ coincides with the initial configuration $\{x_k(0)\}$, but the end-point $x_k(1)$ of the kth string x_k may differ from the initial point $x_k(0)$ of this string, having thus that

$$x_k(1) = Ax_k(0)\,,$$

where A is a permutation of the points which form the initial configuration, denoted with an asterisk, $* \in \mathrm{Conf}_n(\mathbb{R}^2)$.

The elements of the braid group are the homotopy classes of the geometric braids $\{t, x_k(t)\}_{t\in[0,1]}$ whose starting and ending configurations coincide with the initial point $* \in \mathrm{Conf}_n(\mathbb{R}^2)$ to which the fundamental group $\pi_1\left(\mathrm{Conf}_n(\mathbb{R}^2), *\right)$ is associated:

$$\{x_1(0), x_2(0), \ldots, x_n(0)\} \;=\; * \;=\; \{x_1(1), x_2(1), \ldots, x_n(1)\}\,.$$

Example. There is only one (trivial) 1-string braid. Its homotopy class contains the graph of the constant map $x_1(t) = 0$ (for any $t \in I$).

Example. There exist trivial and nontrivial braids with 2 strings (fig. 5.2).

To study the braids, it is useful to understand geometrically the product operation of loops in the fundamental group.

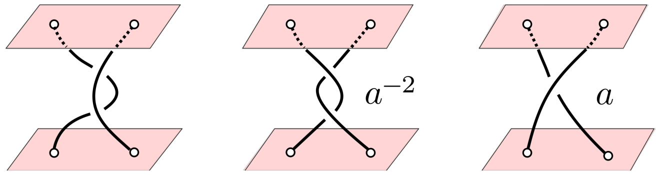


Figure 5.2: A trivial braid with 2 strings (left) and two nontrivial ones (right).

The meaning of the braid multiplication is simply the following one. We put the first n-string braid in the strip $0 \leqslant t \leqslant 1$ of the 3-space $\mathbb{R} \times \mathbb{R}^2$, the second n-string braid in the strip $1 \leqslant t \leqslant 2$, and we obtain the n-string product braid in the strip $0 \leqslant t \leqslant 2$ (which may be reduced to the standard strip $0 \leqslant \widetilde{t} \leqslant 1$ by the transformation $\widetilde{t} = t/2$, if we insist that the loop is a map of the standard segment $I = [0, 1]$).

Traditionally, the drawings are made in such a way that the coordinate t axis in $\mathbb{R}^3$ is vertical, directed downwards. This tradition comes from the usual braids of the young girls, directed down from the head.

Claim. *The group* Br(2) *of braids with two strings is isomorphic to* $\mathbb{Z}$*, its generator being the braid a shown in Fig.* 5.2.
(its strings "making one half-turn", permute the two starting points.)

To prove it, observe that the space of ordered configurations $\mathbb{R}^4 \setminus \Delta_{1,2}$ is diffeomorphic to the product $\mathbb{C} \times (\mathbb{C} \setminus 0)$, and that the permutation acts on this product by sending $(z, w \neq 0)$ to the point $(z, -w)$.

Therefore, the unordered configuration space $\mathrm{Conf}_2(\mathbb{R}^2)$ is diffeomorphic to the manifold $\mathbb{C} \times (\mathbb{C} \setminus 0)$ (with coordinates (z, w^2)). Since the first factor has no importance, and the second one is homotopically contractible to its circle $\mathbb{S}^1$ (defined by the equation $|w^2| = 1$), we find

$$\begin{aligned} \mathrm{Br}(2) &\simeq \pi_1\left(\mathrm{Conf}_2(\mathbb{R}^2), *\right) \simeq \pi_1\big(\mathbb{C} \times (\mathbb{C} \setminus 0), *\big) \simeq \pi_1(\mathbb{C} \setminus 0, *) \\ &\simeq \pi_1(\mathbb{S}^1, *) \simeq \mathbb{Z} \,. \end{aligned}$$

For the braids of three strings the situation is more complicated. There are two generating braids a and b shown in Fig. 5.3 (similar to the generator a of the 2-string braid group Fig. 5.2). These two generators do not commute.

Theorem 1. *The elements a and b of* Fig. 5.3 *verify the relation $aba = bab$ in the braid group* Br(3).

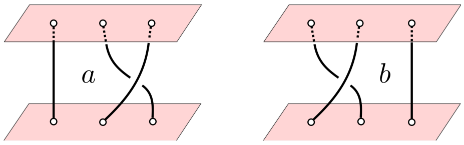


Figure 5.3: Generating braids with 3 strings.

Proof. The necessary homotopy is visible in Fig. 5.4 (where, in the intermediate planes, one can move the $*$-configuration performing the homotopy). $\square$

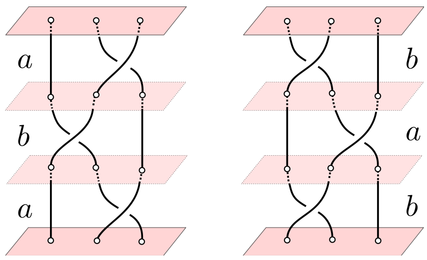


Figure 5.4: Proof of the relation $aba = bab$ in the group of 3-strings braids.

Similarly, for the n-string braids we choose the initial configuration $*$ to be the standard arithmetic progression $\{1, 2, \dots, n\}$ along a line of $\mathbb{R}^2$, defining then the standard braids $\sigma_1, \sigma_2, \dots, \sigma_{n-1}$ of Fig. 5.5. The braid σ_k is similar to the braid a of Fig. 5.2, but here σ_k activates the branches starting at k and $k+1$, by permuting these two starting points at the end moment $t = 1$.

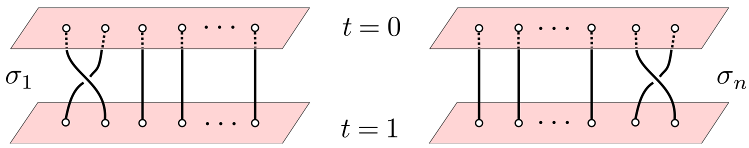


Figure 5.5: The generators σ_k of the group $\mathrm{Br}(n)$.

Theorem 2. *The elements σ_k of the n-string braid group* $\mathrm{Br}(n)$ *verify the relations*

$$\sigma_k\sigma_{k+1}\sigma_k = \sigma_{k+1}\sigma_k\sigma_{k+1} \qquad \textit{for } 1 \leqslant k < n-1 \quad \textit{and}$$
$$\sigma_m\sigma_\ell = \sigma_\ell\sigma_m \qquad \textit{for } |k-\ell| > 1.$$

The first relation is proved by Figure 5.4 (the case $n > 3$ is similar to $n = 3$). The second relation is even more evident geometrically – Fig. 5.6.

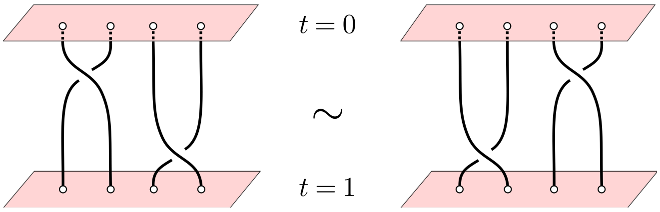


Figure 5.6: Commutativity of σ_1 and σ_3 in Br(4).

In fact, the braid group of n strings is just the group generated by the above generators σ_k, whose defining relations are those of the above theorem.

The fact that the group is generated by the standard braids σ_k is easily proved by the following construction.

Project the branches $x_k(t)$ from $\mathbb{R}^3$ to the 2-plane (t, y), choosing some linear coordinate $y : \mathbb{R}^2 \to \mathbb{R}$ in the plane $\mathbb{R}^2$. Choosing y generically, we may achieve the situation where the projected branches intersect each other only a finite number N of times, transversely and at different time moments $t_1 < t_2 < \cdots < t_N$.

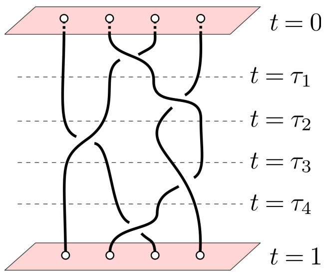


Figure 5.7: Decomposition of a braid in a product of generators.

Decomposing the slice of the 3-space containing the braid, $0 \leqslant t \leqslant 1$, into N slices separated by the planes $t = \tau_j := (t_j + t_{j+1})/2, \quad j < N$,

$$\left(0 \leqslant t \leqslant \frac{t_1 + t_2}{2}\right), \left(\frac{t_1 + t_2}{2} \leqslant t \leqslant \frac{t_2 + t_3}{2}\right), \ \ldots, \ \left(\frac{t_{N-1} + t_N}{2} \leqslant t \leqslant 1\right),$$

the initial braid is represented as the product of N elementary braids (Fig. 5.7). Each of these braids is non-trivial only near the corresponding intersection point of the two projected branches at the time t_k.

This geometric decomposition into N slices presents the initial braid (representative of its homotopy class) as the product of $N+1$ generators (a generator σ_{k_j} or $\sigma_{k_j}^{-1}$ corresponds to the slice of t_j).

All relations of the braid group $\mathrm{Br}(n)$ are simple algebraic corollaries of the standard relations proved above, but the fact that there are no other relations is less simple to prove, and we shall not write its proof here.

5.1.1 Braid group and complement to the swallowtail

The calculation of the braid group may be considered as the calculation of the fundamental group of the complement to an algebraic hypersurface in the space $\mathbb{C}^n$, called “swallowtail”.

Indeed, the configuration space of n unordered points $\{z_1, \dots, z_n\}$ in $\mathbb{C}$ may be considered as the space $\mathbb{C}^n$ of the complex polynomials of degree n,

$$z^n + a_1 z^{n-1} + \cdots + a_n = \prod_{k=1}^{n} (z - z_k)\,,$$

that have no multiple root.

Observe that the ordered set of the n values of the coefficients a_i of each polynomial $z^n + a_1 z^{n-1} + \cdots + a_n$ determines the (unordered) set of its roots. This correspondence is one-to-one and continuous in both directions.

If we allow the possibility of the coincidence of roots (excluding no diagonal), then this correspondence equips the configuration space of n unordered points (which can coincide) with a structure of a complex manifold: that of $\mathbb{C}^n = \{z^n + a_1 z^{n-1} + \cdots + a_n\}$. In such case the *Vieta map*, $V : \mathbb{C}^n \to \mathbb{C}^n$, which sends the n ordered points $(z_1, \dots, z_n)$ to the polynomial

$$\left(a_1 = -(z_1 + \cdots + z_n)\,,\ \ a_2 = z_1 z_2 + \dots,\ \ \cdots,\ \ a_n = (-1)^n \prod_i z_i \right),$$

is an $n!$-fold branched covering of the space of n-th degree monic polynomials, by the ordered configurations of n points in $\mathbb{C} \simeq \mathbb{R}^2$. Its branching locus Σ, which coincides with its set of critical values, is the algebraic hypersurface formed by the polynomials having coincident roots. Hence Σ is the algebraic hypersurface defined by the equation $\Delta(a) = 0$, where $\Delta(a)$ is the discriminant of the polynomial $z^n + a_1 z^{n-1} + \cdots + a_n$.

We conclude that $\mathrm{Conf}_n(\mathbb{R}^2) \approx \mathbb{C}^n \smallsetminus \Sigma$. Therefore, the restriction of the Vieta map to the ordered configurations of n non-coinciding points in $\mathbb{C} \simeq \mathbb{R}^2$

defines a true $n!$-fold covering of the space of n-th degree monic polynomials having no multiple roots.

Example (Complement of the Semi-cubic Parabola). The polynomials of degree three can be reduced to the case $a_1 = 0$ (by a shift on the coordinate z). So for $f(z) = z^3 + az + b$ the vanishing of the discriminant is the condition on the coefficients (a, b) for which the polynomials f and df/dz have a common root: $z^3 + az + b = 0,\ 3z^2 + a = 0$.

For a multiple root z, we get $a = -3z^2,\ b = -z^3 + 3z^3,$ whence the set of polynomials with multiple roots are described by the equation

$$\left(\frac{a}{3}\right)^3 + \left(\frac{b}{2}\right)^2 = 0\ . \tag{1}$$

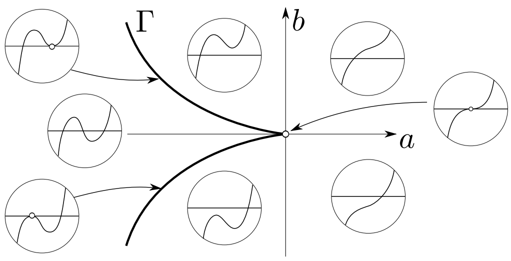


Figure 5.8: The curve Γ of cubic polynomials of discriminant zero as the root bifurcation diagram.

This semicubic curve Γ (Fig. 5.8) is the set of polynomials of discriminant zero in our family. Returning to the general polynomials of degree 3, we obtain the real diffeomorphism

$$\mathrm{Conf}_3(\mathbb{R}^2) \approx \mathbb{C} \times (\mathbb{C}^2 \smallsetminus \Gamma)\ ,$$

whence

$$\mathrm{Br}(3) = \pi_1\big(\mathrm{Conf}_3(\mathbb{R}^2), *\big) \approx \pi_1\big(\mathbb{C}^2 \smallsetminus \Gamma, *\big)\ . \tag{2}$$

Example (Complement of the Swallowtail). One can reduce the study of the polynomials of degree n to those of trace zero, $a_1 = 0$, by a choice of the origin of the z axis. Then we have to study the complement $\mathbb{C}^{n-1} \smallsetminus \widetilde{\Sigma}^{n-2}$ to the algebraic hypersurface $\widetilde{\Sigma}^{n-2}$ formed by the zero-trace polynomials having multiple roots. This surface is called (complex) *swallowtail*.

Thus, the 3-string and 4-string braid groups are the respective fundamental groups of the complement to the complex semicubic algebraic curve in the complex plane $\mathbb{C}^2$ and to the swallowtail surface in $\mathbb{C}^3$.

Similarly, the group $\mathrm{Br}(n+1)$ is the fundamental group of the complement to the complex $(n-1)$-dimensional swallowtail hypersurface in $\mathbb{C}^n$.

Classifying Space. A space E is called *classifying space* for its fundamental group $\pi_1(E)$ if all higher homotopy groups of this space are trivial.

Since we have shown that all homotopy groups π_k of $Conf_n(\mathbb{R}^2)$ are trivial (p. 102), we deduce from the above problem the following general result.

Corollary. *The complement to the semicubic parabola in* $\mathbb{C}^2$ *is the classifying space of the braid group* $\mathrm{Br}(3)$. *The complement to the higher dimensional swallowtails are the classifying spaces of the braid groups of* n *strings* $\mathrm{Br}(n)$.

The topological invariants of the classifying space provide the corresponding "topological invariants" of the groups ("cohomology groups", etc).

The real swallowtail.

The real swallowtail surface was first studied by Kronecker (for some number-theoretical reasons) as the space of real polynomials of degree 4 having discriminant zero.

This Kronecker surface is shown in figure 5.9. It subdivides the real 3-space of the polynomials with real coefficients x^4+ax^2+bx+c into three parts, which consists of those polynomials that have 4, 2 or 0 real roots.

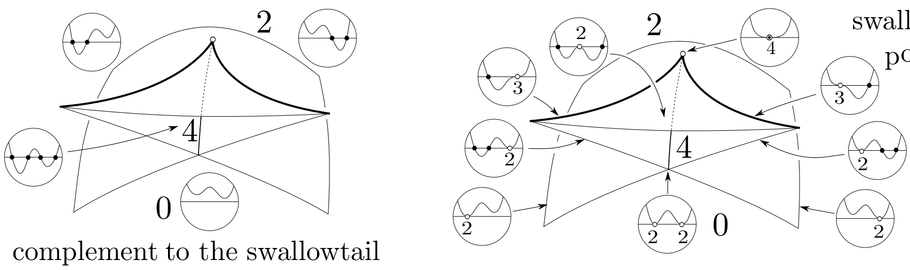


Figure 5.9: The swallowtail surface as the set of polynomials x^4+ax^2+bx+c of discriminant zero in the space with coordinates (a,b,c).

It is a useful exercise to discover which parts of the complement to the swallowtail in Fig. 5.9 correspond to which number of roots. Verify that Fig. 5.9 is correct.

One general principle, helping here, is the philosophy of the instability and vulnerability of all good things. Guessing that it is good to have many real roots, we guess that the 4 real roots part is the smallest of the three ones (in the neighbourhood of the origin, where all the roots are equal to zero). So, in Fig. 5.9, the polynomials with four real roots form the interior part of the pyramid, whose three edges are formed by the semicubic cuspidal edge upstairs and the selfintersection line downstairs. (Similarly, in Fig. 5.8 the polynomials with three real roots live in the angular domain bounded by the curve Γ.)

The neighbouring (upper) domain contains the polynomials with two real roots, and the part formed by the polynomials with no real roots is below the swallowtail surface.

5.2 Zariski Theorem 1

The fundamental group of the complement to an arbitrary algebraic hypersurface Σ is described by three theorems of Zariski that we shall explain.

To study the fundamental group $\pi_1(\mathbb{C}^n \smallsetminus \Sigma, *)$, we start by projecting the complex algebraic hypersurface Σ of $\mathbb{C}^n$ along the parallel fibres of a linear projection $p : \mathbb{C}^n \to \mathbb{C}^{n-1}$.

The chosen projection must be generic, avoiding exceptional directions. For example, for the semicubic parabola $\Gamma \subset \mathbb{C}^2$ of page 139 (Fig. 5.8) the projection $p(a, b) = a$ is convenient; for the swallowtail surface $\widetilde{\Sigma}^{n-2} \subset \mathbb{C}^{n-1}$ (in the space of zero-trace polynomials $a_1 \equiv 0$) a convenient projection is $p(a_2, \ldots, a_n) = (a_2, \ldots, a_{n-1})$. The fibre $\mathbb{C}_*$ that contains the basic point $*$ also should be generic; these genericity conditions are discussed below.

Denote by $\alpha_1, \ldots, \alpha_\mu$ the (different) intersection points of the chosen fibre $\mathbb{C}_*$ with the hypersurface Σ (fig. 5.10).

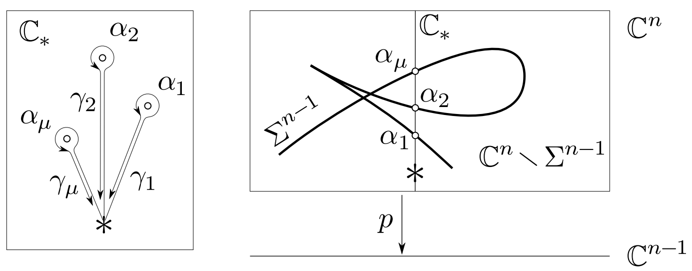


Figure 5.10: Zariski generators γ_k of the fundamental group of the complement to an algebraic hypersurface in $\mathbb{C}^n$.

The complement $\mathbb{C}_* \smallsetminus \Sigma$ to these μ points of the complex line $\mathbb{C}_*$ (the fibre) is included in the domain $\mathbb{C}^n \smallsetminus \Sigma$.

The fundamental group of the complement to Σ in the fibre,

$$\pi_1(\mathbb{C}_* \smallsetminus \Sigma, *) = F_\mu,$$

is the free group generated by the μ simple loops $\gamma_1, \dots, \gamma_\mu$, encircling the μ excluded points (Fig. 5.10). The inclusion $(\mathbb{C}_* \setminus \Sigma) \to (\mathbb{C}^n \setminus \Sigma)$ sends this free group into the fundamental group $\pi_1(\mathbb{C}^n \setminus \Sigma, *)$ that we wish to study.

Zariski Theorem 1. *The image of the free group F_μ covers the fundamental group $\pi_1(\mathbb{C}^n \setminus \Sigma, *)$ of the complement to the hypersurface Σ. That is, the desired fundamental group is generated by the μ simple loops $\gamma_1, \dots, \gamma_\mu$, living in the chosen generic fibre $\mathbb{C}_*$ of our generic projection.*

These μ generators may have non-trivial relations, and the desired group is in general different from F_μ. We shall describe these relations below.

Example (The "circle"). Consider the curve $\Sigma : \{x^2 + y^2 = 1\}$ in $\mathbb{C}^2$, the projection $p(x, y) = x$ and the basic point $* = (0, -2)$, which is in general position on the generic fibre $\mathbb{C}_* : \{x = 0\}$ (parametrised by the y coordinate).

There are two points of the curve Σ on the fibre line $\mathbb{C}_*$, say α_1 at $y = -1$, and α_2 at $y = +1$ – Fig. 5.11.

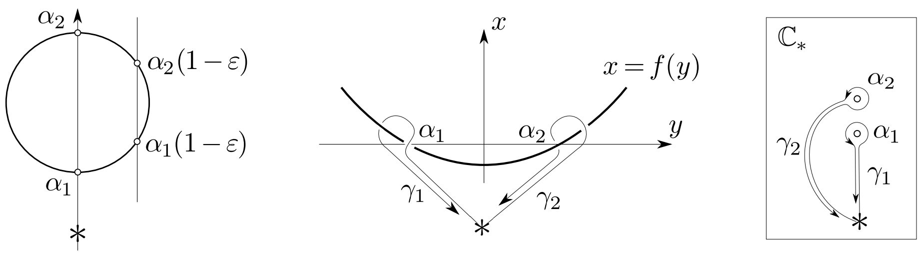


Figure 5.11: Calculation of the fundamental group of the complement to the "circle" in the complex plane.

Zariski Theorem 1 claims that the loops γ_1 and γ_2 generate the fundamental group of the complement to the circle. But this group is the cyclic group

$$\pi_1\big(\mathbb{C}^2 \setminus \{x^2 + y^2 = 1\}, *\big) \approx \mathbb{Z}$$

because its "generators" γ_1 and γ_2 verify in it the relation $\gamma_1 = \gamma_2$.

Indeed, moving the point $*$ and the fibre $x = 0$ towards the line $x = 1$ by a homotopy, we arrive at $x = 1 - \varepsilon$ to the situation of two almost colliding points of intersection $\alpha_k(x)$ of the circle with the fibre $\mathbb{C}_*$. At this situation, near the collision point $x = 1$, $y = 0$, we can write the local equation of the curve in the form $x = f(y)$ for a smooth function f (by the implicit function theorem). Hence the fundamental group of the local complement is

$$\begin{aligned}\pi_1\big(\mathbb{C}^2 \smallsetminus \{x = f(y)\}, *\big) \simeq \pi_1\big(\mathbb{C} \times (\mathbb{C} \smallsetminus 0), *\big) &\simeq \pi_1(\mathbb{C} \smallsetminus 0, *)\\ \simeq \pi_1(\mathbb{S}^1, *) \simeq \mathbb{Z}\,.&\end{aligned}$$

For instance, we obtain this way a homotopy between the loops γ_1 and γ_2, which, therefore, represent the same element of the fundamental group of the complement to the complex circle in the complex plane.

Example ("Degenerate circle"). For the curve in $\mathbb{C}^2$ defined by the equation $x^2 + y^2 = 0$, the above reasoning provides the fundamental group of its complement: $\pi_1(\mathbb{C}^2 \smallsetminus \{x^2 + y^2 = 0\}, *) \approx \mathbb{Z}^2$. It is the commutative group with generators γ_1 and γ_2 satisfying the defining relation $\gamma_1\gamma_2 = \gamma_2\gamma_1$.

Indeed, using the linear complex coordinate system

$$(z = x + iy,\ w = x - iy)\,,$$

we reduce the equation of the degenerate curve to the form $zw = 0$, whence

$$(\mathbb{C}^2 \smallsetminus \{x^2 + y^2 = 0\}) \approx (\mathbb{C}^2 \smallsetminus \{zw = 0\}) \approx (\mathbb{C} \smallsetminus 0) \times (\mathbb{C} \smallsetminus 0)\,,$$

and therefore we get the isomorphism of the desired fundamental group with the fundamental group of the two-dimensional torus,

$$\pi_1(\mathbb{C}^2 \smallsetminus \{x^2 + y^2 = 0\}) \approx \pi_1(\mathbb{S}^1 \times \mathbb{S}^1) \approx \mathbb{Z}^2\,.$$

Necessity of a generic projection. The choice of a nongeneric projection might produce a different topological situation, while for all generic choices the situation will be the same.

Example. For the projection $(z, w) \mapsto z$ along the asymptotic direction of the hyperbola of the preceding example, the generic fibre $\mathbb{C}_* : \{z = 1\}$ has only one intersection point with the degenerate curve $zw = 0$ (and with the hyperbola $zw = 1$). In this case, Zariski Theorem 1 fails because the loop obtained from the intersection point does not generate the fundamental group $\mathbb{Z}^2$ of the complement to the degenerate curve.

Necessity of a generic fibre. Even in the case of a generic projection direction, a nongeneric choice of the basic fibre $\mathbb{C}_*$ might produce an example in which the fundamental group is not generated by the loops γ_k.

Example. For the projection that sends (z, w) to $t = z + w$, one should not take the degenerate fibre $\mathbb{C}_* : (t = 0)$ to study the fundamental group

$$\pi_1\big(\mathbb{C}^2 \setminus \{zw = 0\}, *\big) \approx \mathbb{Z}^2$$

because this fibre would provide only one intersection point α_k and only one loop γ_k.

Similarly, in the case of the hyperbola $zw = 1$ (Fig. 5.12), to study the fundamental group

$$\pi_1(\mathbb{C}^2 \setminus \{zw = 1\}) \approx \mathbb{Z}$$

one should not take the degenerate fibre $\mathbb{C}_* : \{t = 1\}$ because the corresponding loop γ would be homotopic to 0 in $\mathbb{C}^2 \setminus \{zw = 1\}$.

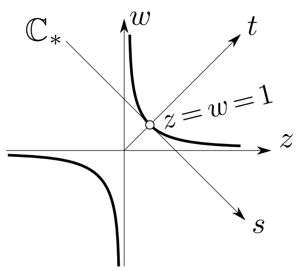


Figure 5.12: A non generic fibre $\mathbb{C}_*$ of a generic projection (parallel to it).

This follows from the implicit function theorem, which provides the local equation $t = f(s)$ of this hyperbola ; we use the local coordinate system $(t = z + w, s = z - w)$ near the point $(z = w = 1)$ of the hyperbola.

5.3 Italian principle

The above examples show that in Zariski Theorem 1 one must not forget the genericity conditions on the projection p and on the fibre $\mathbb{C}_*$.

The following general statement is true in many different problems and is known as "the Italian principle in complex algebraic geometry".

Suppose the complex algebraic geometry objects we are studying depend on complex parameters and consider a smooth connected complex algebraic variety of these objects. For example, the space $\mathbb{C}^n$ of complex polynomials of degree n,

$$f_a(z) = z^n + a_1 z^{n-1} + \cdots + a_n \tag{3}$$

in one complex variable $z \in \mathbb{C}$, or the variety of all the complex projective algebraic curves of fixed degree n in the complex projective plane $\mathbb{C}\mathrm{P}^2$.

Italian Principle. *For the generic choices the topology of the situation will not depend on these choices.*

The statement that all generic objects of such a family are topologically similar, follows from a simple (implicit function theorem type) reasoning. The degenerate objects (like the polynomials having a multiple root, or the curves having singularities) form a *complex* algebraic subvariety in our family. For example, in the case (3) this subvariety is defined by the algebraic equation obtained from the condition that the polynomial discriminant $\Delta(a_1, a_2, \ldots, a_n)$ is equal to zero; and, in the case of the curves, by a similar equation on the coefficients of the equations which define those curves.

For a nondegenerate object the topological situation is stable. The number of roots of a polynomial (3) with no multiple roots remains constant under a small variation of the coefficients (see Fig. 5.13).

The small deformations of the coefficients of the equation of a smooth (complex projective) curve of degree n provide deformed curves, homeomorphic to the original one (unlike the deformation of the singular curve $zw = 0$ to the smooth curve $zw = \varepsilon$).

If the degeneracy condition is defined by one complex equation, the base of the "Italian principle" is as follows:
The single complex equation "discriminant equals zero" means two real conditions, and hence the variety of singular objects (of complex polynomials with multiple roots or algebraic curves with singular points) *has real codimension at least two in the whole manifold of parameter values.*

It follows that *the family of nonsingular objects is connected*: one can reach each nonsingular object from any other nonsingular object by a smooth path in the complement to the hypersurface of singular objects.

Since the topological properties of the non-singular objects are locally constant, we conclude that they are also globally constant because all nondegenerate objects have the same topological type (even if the parameter values are very different from those of the initial generic object, to which we compare the others). Let us see the simplest examples.

5.3.1 Fundamental Theorem of Algebra

Fundamental Theorem of Algebra. *Every complex polynomial* (3) *of* *degree n with no multiple roots has n distinct complex roots.*

Proof. According to the above principle, it suffices to present one such polynomial, say

$$f = (z-1)(z-2)\dots(z-n)\ .$$

Its roots are evidently just the numbers $\{1, 2, \dots, n\}$. Hence, by the "Italian argument" any complex polynomial (3) with no multiple roots has exactly n complex roots. So, the "fundamental theorem of algebra" is just a simple geometric remark (Fig. 5.13). $\square$

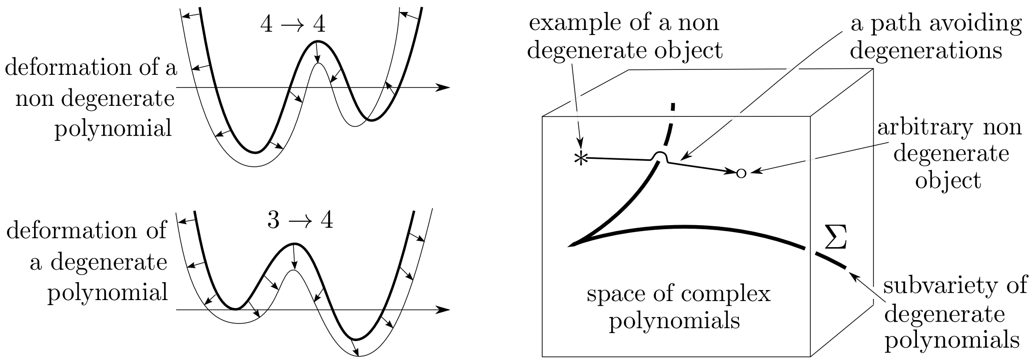


Figure 5.13: Geometric proof of the fundamental theorem of algebra.

5.3.2 Homeomorphic Algebraic Curves

Homeomorphic Algebraic Curves. *All nonsingular complex algebraic curves of degree n in the complex projective plane are homeomorphic.*

Proof. It follows from the general principle because the non-singular curve remains homeomorphic to itself under any sufficiently small variation of the coefficients of the equation, and the set of all curves of degree n is the complex projective space $\mathbb{C}\mathrm{P}^{N(n)}$, which is a connected smooth complex manifold. $\square$

The computation of the dimension $N(n)$ in the above proof is a useful exercise: $N(1) = 2$, $N(2) = 5$, $N(3) = 9$ – one should count the integral points on the Newton simplex $u + v + w = n$, $u \geqslant 0$, $v \geqslant 0$, $w \geqslant 0$ of the "Newton parallelogram space" $\mathbb{Z}^3$ of the monomials $x^u y^v z^w$ – Fig. 5.14. .

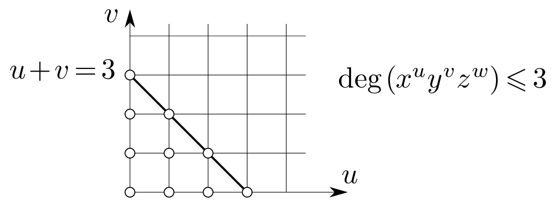


Figure 5.14: Newton parallelogram space (of the degrees of the monomials).

5.3.3 Genus Formula

Genus of a Curve. A complex curve in the complex projective plane, regarded as a real object, is a 2-dimensional compact real surface (see p.6). The genus of this surface is called the *genus of the complex curve*.

Problem. Find the genus (topological type) of all algebraic plane curves.

Solution. By the "Italian principle", it suffices to study topologically one example of a non-singular curve of degree n.

As the required example, it is useful to take first a very degenerate curve consisting of n generic straight lines which intersect each other at

$$(n-1)+(n-2)+\cdots+1 = n(n-1)/2$$

places – Fig. 5.15.

Each of these complex projective lines is diffeomorphic to $\mathbb{S}^2(=\mathbb{C}\mathrm{P}^1)$. But the union of these n spheres is not smooth, since it contains all the $n(n-1)/2$ transverse intersections, one for each pair of spheres.

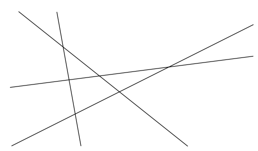


Figure 5.15: A completely degenerate curve of degree n consists of n straight lines.

Consider the equation $f=0$ of our degenerate curve: the polynomial f of degree n is the product of the n polynomials of the degree 1, which define the lines. Now, replace the equation $f=0$ by its variation $f=\varepsilon$ (we write the equations in affine coordinates of the plane, supposing that all intersection points of our lines are finite, and the lines are not parallel).

To understand what happens near the intersection point, under the deformation $0 \mapsto \varepsilon$ to the singular curve, it suffices to study the simplest example (to which the general case can be reduced by a local coordinate change) : the transition from the two intersecting lines $zw = 0$ to the hyperbola $zw = \varepsilon$.

The topological smoothening of a complex intersection point is the following operation. We delete a small disc, centred at the intersection point, from each intersecting sphere; then we join two such intersecting spheres by gluening to them a cylindrical tube along the boundaries of their small holes – Fig. 5.16.

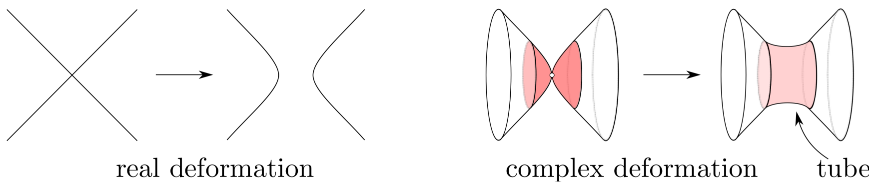


Figure 5.16: Topological structure of the standard smoothening operation of the transverse intersection point of two curves in the real and in the complex domains.

This topological description of the transition from the complex curve $zw = 0$ to the curve $zw = \varepsilon$ follows, for instance, from the preceding study of the topology of the complex hyperbola (and of the complex circle). Taking on the hyperbola the coordinate $z \neq 0$, we see that it is $\mathbb{R}$-diffeomorphic to the cylinder $\{z \neq 0\} = (\mathbb{C} \setminus 0) \sim (\mathbb{S}^1 \times \mathbb{R})$.

Therefore, *the complex circle* $\{x^2 + y^2 = 1\}$ *in the complex plane* $\mathbb{C}^2$ *is* $\mathbb{R}$-*diffeomorphic to the cylinder* $\mathbb{S}^1 \times \mathbb{R}$.

Returning to our degenerate singular curve of degree n, we have to add $n(n-1)/2$ cylinders at its $n(n-1)/2$ points of intersection (of the spheres forming it). To understand the topological effect of this operation, we first delete the $n(n-1)$ small neighbouring discs at all the intersections, and then, choosing a "first" sphere, we add the $n-1$ cylindrical tubes, joining this "first" sphere to all the others – Fig. 5.17.

The result is a surface homeomorphic (or $\mathbb{R}$-diffeomorphic) to the sphere with $n(n-1) - 2(n-1) = (n-1)(n-2)$ remaining holes.

Adding the $(n-1)(n-2)/2$ joining tubes at the remaining pairs of holes (provided by the pairs of intersecting spheres, different from the "first" chosen one), we complete the smoothening operation. The resulting surface is a

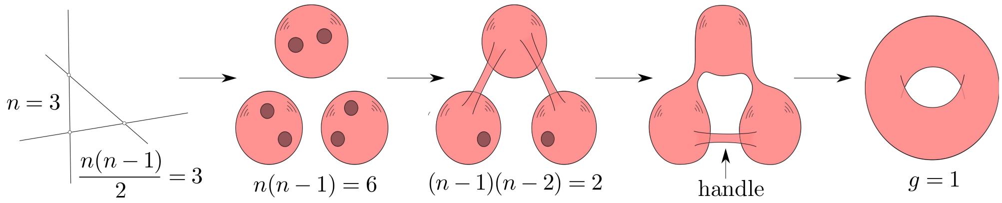


Figure 5.17: Proof of the genus formula for the deformed singular curve.

sphere with

$$g = \frac{(n-1)(n-2)}{2} \tag{4}$$

handles, one for each tube attached at the last step of the operation.

This *genus formula* of Riemann-Hurwitz, which gives the topological description of the non-singular algebraic curves of degree n in the complex projective plane, is a simple corollary of the general "Italian principle of algebraic geometry", for which it is enough to study a particular example.

Remark. The genus formula (4) tell us that not all compact surfaces are realisable as a non-singular algebraic curve in the complex projective plane. For example, there is no such curve with genus $g = 2,\ 4,\ 5,\ 7,\ 8$ or 9.

5.4 On Real Algebraic Geometry

As we mentioned above, the "Italian principle" in *complex* algebraic geometry is based on the fact that *the complex variety of degenerate objects has real codimension* 2*, making all non-singular objects attainable from one another by continuous paths in the domain of non-singular objects.*

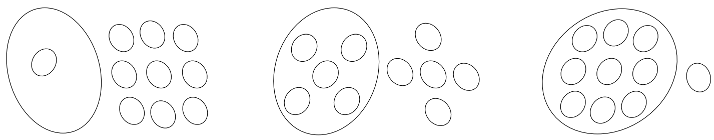
In *real* algebraic geometry the varieties formed by the degenerate objects are of real codimension 1, separating the domain of non-degenerate objects into different components. Therefore, to join some pairs of non-degenerate objects by a continuous path, it is unavoidable to cross these varieties.

This explains the difference between the numbers of real roots of different nondegenerate polynomials of fixed degree (like the 3 domains of the complement to the swallowtail in Fig. 5.9), and the difficulty to computate these numbers of real roots for non-degenerate real polynomials.

Topological Types of Real Curves. We have found above the topological types of all generic complex algebraic curves of the projective plane for any degree n. The corresponding classification of topological types of real curves of degree n in the real projective plane is a fundamental problem of mathematics which is still open even for algebraic curves of degree 8.

Harnack proved that the number of ovals (closed component, diffeomorphic to a circle) of a curve of genus g in the real projective plane can't exceed the number $g+1$ (being one for a conic and two for an elliptic curve). The curves having these maximal number of ovals, $g+1$, are called M-curves.

Hilbert included in his 16 th problem the question on the possible topological configurations of these ovals. For M-curves of degree $n=6$ Hilbert claimed that there exist only two configurations of the $g+1=11$ ovals. However, D. A. Gudkov proved that there are three configurations – Fig. 5.18.



$\chi(\{(x,y) : f(x,y) \geqslant 0\}) \equiv k^2 \pmod 8$
for a curve $f=0$ of degree $n=2k$
having $(n-1)(n-2)/2+1$ ovals.

Figure 5.18: The three Gudkov topological configurations of the $g+1=11$ ovals of a real plane algebraic M-curve of degree $n=6$. The middle curve is the counterexample to Hilbert's claim.

The study of Gudkov's configurations is related to the topology of 4-dimensional smooth manifolds and to quantum field theory, but we must resist the temptation to describe these nice theories in this textbook.

Anyway, real ($\mathbb{R}$) problems are too difficult for the present-day algebraic geometers, preferring complex ($\mathbb{C}$) theories with their Italian principle. It is interesting that computer science has still contributed nothing to the basic problem on the topological shapes of the real algebraic curves.

5.5 Rational Curves

The Riemann-Hurwitz formula expresses that for degrees $n=1$ and 2 the number of handles is $g=0$. That is, the complex projective curves of degrees 1 and 2 (lines and conics in $\mathbb{C}\mathrm{P}^2$) are real diffeomorphic to the sphere $\mathbb{S}^2$.

This remark is helpful to memorise the Riemann-Hurwitz formula. It also explains the reasons for the simple Newton integration of the square roots of polynomials of degree 2 (including the integrals like $\int dx/\sqrt{x^2+ax+b}$).

Namely, the algebraic curves of genus $g = 0$ (which are $\mathbb{R}$-diffeomorphic to sphere $\mathbb{S}^2 \approx \mathbb{C}\mathrm{P}^1$) are called *rational curves*, since they have rational parametrisations

$$x = p(t) \ , \quad y = q(t)$$

for some rational functions p and q (of the affine coordinate t on the parametrising complex projective line, sent by (p, q) to the algebraic curve in the complex affine plane $\{x, y\}$, that is studied). For this reason they are $\mathbb{C}$-diffeomorphic to $\mathbb{C}\mathrm{P}^1 = \mathbb{S}^2$.

For the standard circle $x^2 + y^2 = 1$ such a rational parametrisation is provided by the classical trick, describing all the Egyptian rectangular triangles with integer lengths of the sides, like $3^2 + 4^2 = 5^2$.

One starts at some point of the curve (say, $x = -1$, $y = 0$ on our circle, fig. 5.19).

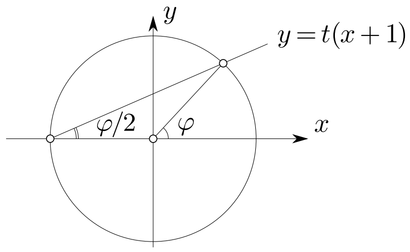


Figure 5.19: The rational parametrisation of the circle by the tangent of the half-angle parameter.

The straight line of slope t, containing this point, is defined by the equation $y = t(x+1)$. Putting this expression in the equation, we find a quadratic equation for the coordinate x of the intersection point of the line with the curve.

But one root of this equation is provided by the chosen starting point $(x = -1)$. Therefore, the second root is given by the Vieta formula, providing a rational function $x(t)$ (its explicit computation is a useful exercise).

This calculation provides the rational parametrisation of the circle:

$$x = \frac{1 - t^2}{1 + t^2} \ , \qquad y = \frac{2t}{1 + t^2},$$

called *half-angle tangent formula*, since $t = \mathrm{tg}(\varphi/2)$ for $(x = \cos\varphi, y = \sin\varphi)$.

To get the Egyptian triangles $X^2 + Y^2 = Z^2$, observe that the number $t = y/(x + 1)$ is rational if x and y are rational numbers; and conversely, for any rational number $t = u/v$ we have obtained the explicit coordinates of the rational point $(x(t), y(t))$ on the circle:

$$x = \frac{u^2 - v^2}{u^2 + v^2}\,, \qquad y = \frac{2uv}{u^2 + v^2}\,.$$

Therefore, *the Egyptian triple is given by*

$$X = u^2 - v^2\,, \qquad Y = 2uv\,, \qquad Z = u^2 + v^2$$

with integers u and v. Thus, the values $(u = 2, v = 1)$ provide the triple $(X = 3,\ Y = 4,\ Z = 5)$, while the values $(u = 3, v = 2)$ generate the triple $(X = 5,\ Y = 12,\ Z = 13)$; indeed, $5^2 + 12^2 = 13^2$.

On the other side, for any rational algebraic curve $H(x, y) = 0$ and any rational function $R(x, y)$, the integral

$$I = \int_{H(x,y)=0} R(x, y)\ dx$$

is computable in elementary functions.

Such integrals (for arbitrary algebraic curves $H = 0$ and rational functions R) are called *Abelian integrals*, since Abel invented and studied the Riemann surfaces, trying to investigate these integrals.

To compute the Abelian integral along a rational curve, using the rational parametrisation $x = p(t)$, $y = q(t)$, we get the expression

$$I = \int R\big(p(t), q(t)\big)\ p'(t)\ dt\ .$$

It remains, therefore, to integrate a rational function of the parameter t.

5.6 Elliptic curves

For the quadratic polynomials H, the above trick (providing the Egyptian triples) leads to the so-called "Euler substitutions" for the integrals involving square roots. But they are available only for the *rational curves* (genus $g = 0$). If the topological structure of the complex curve defined by the equation $H(x, y) = 0$ is different from that of the sphere, that is, it has $g > 0$ handles, then the generic Abelian integrals along such a curve are not elementary functions (the exceptional function 0 is easy to integrate).

Example. For a generic cubic curve ($n = 3$ in the preceding theory) the Riemann-Hurwitz formula provides the genus $g = 1$, the complex curve is therefore $\mathbb{R}$-diffeomorphic to the torus.

The algebraic curves of genus $g = 1$ are called *elliptic curves*, since the length of an ellipse is provided by an Abelian integral along such a curve.

A torical holomorphic curve is also obtainable from the affine complex line $\mathbb{C}$ – quotienting $\mathbb{R}^2 \approx \mathbb{C}$ by the commutative group $\mathbb{Z}^2$ of its translations along two $\mathbb{R}$-independent vectors ω_1 and ω_2 – Fig. 5.20. All such curves are algebraic elliptic curves, but their algebraicity is not so evident.

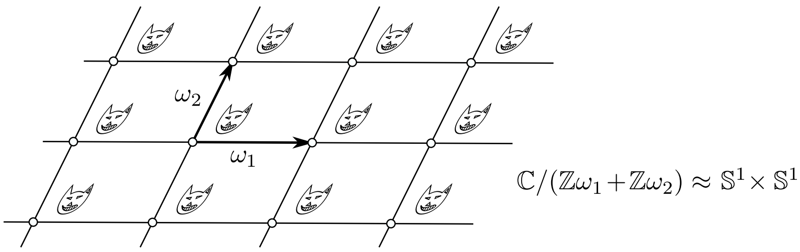


Figure 5.20: The complex holomorphic curve, $\mathbb{R}$-diffeomorphic to the torus $\mathbb{S}^1 \times \mathbb{S}^1$.

Nice examples of elliptic curves are provided by the energy preservation law for the Newton equation with a potential U of degree 3 or 4 :

$$\frac{y^2}{2} + U(x) = E \ , \qquad \frac{d^2x}{dt^2} = F(x) \ ,$$

where the potential energy U is related to the force field F by the defining equation $F = -\partial U/\partial x$ and where the kinetic energy is quadratic with respect to the velocity $y = dx/dt$. The time t of the motion along the curve of fixed constant energy E in the phase plane with coordinate x and y is provided by the Abelian integral :

$$\frac{dx}{dt} = y \ , \qquad dt = \frac{dx}{y} = \frac{dx}{\sqrt{2(E - U(x))}} \ .$$

For a polynomial potential energy U, of degree 2, the curve of constant energy is rational, and hence the integral can be calculated in elementary functions. This integration provides the description of the (harmonic) linear oscillations in terms of trigonometric functions.

For the cases of the degrees 3 or 4, the constant energy curve is elliptic ($g = 1$), and the corresponding Abelian elliptic integral cannot be expressed in terms of elementary functions. This is a new transcendental function – one of the most important functions in mathematics (next after the polynomials, exponentials and trigonometric functions).

These elliptic integrals may be reduced to the following "Weierstrass normal form" :

$$t(X) = \int^X \frac{dx}{\sqrt{x^3 + ax + b}} \, . \tag{5}$$

The inverse function $X(t)$ is a doubly-periodic function (with two complex periods ω_1 and ω_2), meromorphic in the whole affine line $X \in \mathbb{C}$ (Fig. 5.21) :

$$t(X + \omega_1) \equiv t(X + \omega_2) \equiv t(X).$$

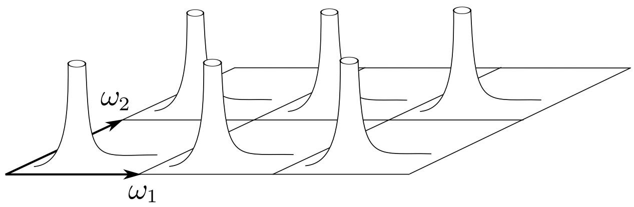


Figure 5.21: A doubly-periodic meromorphic function with one pole on the torus.

Three problems. To understand the usefulness of the preceding geometric theory for some practical questions of nonlinear oscillation theory, think on the following three problems.

PROBLEM. Suppose the potential U is a polynomial of degree 4 with two wells A and B – Fig. 5.22. Consider the following periodic motions with equal total constant energy E : that in the deeper well A and that in the less profound well B. *Is the period of the motion A, in the deeper well, bigger or smaller than the period of the motion B, in the other one*?

ANSWER. These periods are equal.
Though the path A in the deeper well is longer, its velocity is also larger than for B. The torical topology of the complex constant energy curve is responsible of the equality of both periods. This follows from the next problem.

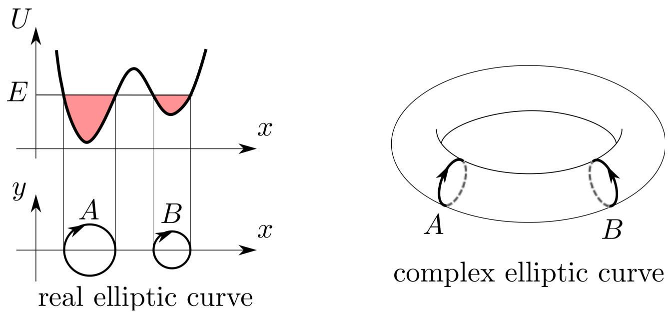


Figure 5.22: Topological proof of the equality of periods on A and B.

Of course, this answer can be also obtained by the transformations of the integrals, whose calculation would, however, provide no understanding of the nature of things.

PROBLEM. Prove that *the time differential form* dx/y *of formula* (5) *has no singularities along the complex elliptic curve in the complex projective plane* (that is, along the whole torus surface).

SOLUTION. The singularities could arise at two places: 1) near the points where $y = 0$; 2) at the infinitely far points of the elliptic curve.

To see that there is no singularity, in the first case one should not choose x as the local coordinate along the constant energy curve; while in the second case one should not use the (x, y) coordinate system, but $(1/x$ and $y/x)$.

Thus, the periods on the two wells are the integrals of the smooth form dt along two different closed real curves A and B on the torus $\mathbb{C}/(\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2)$ (which is $\mathbb{C}$-diffeomorphic to the curve we are studying).

The answer to the previous problem follows from the homology between the two meridians of the torus that represent the two real periodic motions A and B in both wells (these two disjoint real curves are part of the complex elliptic curve, which is connected and $\mathbb{R}$-diffeomorphic to the torus).

PROBLEM. Prove that *neither the elliptic integral* $t = I(X)$ (labeled (5) of p. 154) *nor the inverse doubly-periodic meromorphic function* $X = F(t)$, *are topologically equivalent to any elementary function.*

The reason for the impossibility to express these integrals in terms of elementary functions is topological, as in many other impossibility results. Some topological impossibility theorems are discussed on p. 509. In Chapter 13, we give a topological proof of the impossibility to solve by radicals the algebraic equations of degree 5 or higher.

5.7 Modulus Parametrising the Elliptic Curves

In our discussion about the "Italian principle" we have mentioned the preservation of the "topological properties" of the non-singular objects under their small perturbations.

In the case where the objects were the roots of polynomials, the preserved "topological property" was simply the number of roots.

In the algebraic curves example we have discussed the preservation of the genus g, and the homeomorphic deformation of the manifold under the variation. These homeomorphisms, generated by the implicit function theorem, can be chosen to be even $\mathbb{R}$-diffeomorphisms.

But since our algebraic curves are also complex manifolds (of dimension one) arises the question on whether the deformed curve is complex diffeomorphic to the original one.

In the case of the rational curves ($g = 0$), these complex diffeomorphisms exist. This is a non-trivial theorem of the theory of holomorphic functions, known as "the Riemann theorem on the uniqueness of the holomorphic structure of the sphere".

In the case of curves of higher genus ($g > 0$), there is no such a theorem. Already the elliptic curves

$$\mathbb{C}/(\omega_1\mathbb{Z} + \omega_2\mathbb{Z}) \quad (\mathrm{Im}\ (\omega_2/\omega_1) \neq 0),$$

being all homeomorphic and $\mathbb{R}$-diffeomorphic to the standard real torus $\mathbb{S}^1 \times \mathbb{S}^1$, are, in general, holomorphically different between them: there is generically no holomorphic map between two such curves.

Modulus. *The ratio $\lambda = \omega_2/\omega_1$ is a "modulus"*: a holomorphic invariant that parametrises the different holomorphic types of elliptic curves.

This result follows from the fact that such a holomorphic diffeomorphism would generate a holomorphic diffeomorphism between the covering planes, $\mathbb{C} \to \mathbb{C}$, which should be an affine map (by another classical theorem of the theory of holomorphic functions).

This description of the holomorphic maps provides the invariance of the lattice generated by the periods $\{\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2\}$ (up to a common complex scalar factor). The invariance of $\lambda = \omega_2/\omega_1$ would follow, if the basis of the lattice were chosen. Taking into account the possibility of replacing the basic vectors by their integer linear combinations (of determinant 1), we get the invariance of λ modulo the action of the group of unimodular matrices:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}) : \lambda \mapsto \frac{a\lambda + b}{c\lambda + d}\,, \quad ad - bc = 1\ .$$

Choosing hypothetically the sign of the basic vector, we may suppose $\mathrm{Im}\,\lambda > 0$, and thus the description of the space of elliptic curves is provided by the variety

$$(\mathbb{C}^+ = \{\lambda : \mathrm{Im}\,\lambda > 0\})\,/\,\mathrm{SL}(2,\mathbb{Z}).$$

This complex one-dimensional variety is in fact holomorphically equivalent to the sphere $\mathbb{S}^2 = \mathbb{C}\mathrm{P}^1$ with 3 points deleted (one usually deletes the values 0, 1 and ∞ of some affine coordinate of $\mathbb{C}\mathrm{P}^1$).

For the curves of genus $g > 1$, the space which parametrises their holomorphic types is a complex manifold of complex dimension $3g - 3$.

It is a nice problem to find the way to calculate this dimension.

5.8 Zariski Theorems 2 and 3

Returning to the Zariski description of the fundamental groups of the complements of algebraic hypersurfaces, we shall formulate two more theorems.

Similarly to Zariski Theorem 1, we start by considering an algebraic hypersurface $\Sigma \subset \mathbb{C}^n$, a generic projection $p : \mathbb{C}^n \to \mathbb{C}^{n-1}$ and generic fibre $\mathbb{C}_*$ of p. But, in this case, we shall also take a generic complex 2-plane $\mathbb{C}^2 \subset \mathbb{C}^n$ containing the chosen fibre $\mathbb{C}_*$.

The complement to Σ in this plane is included into the complement to Σ in $\mathbb{C}^n$. Then the fundamental group of the complement to the algebraic curve $\Sigma \cap \mathbb{C}^2$ in $\mathbb{C}^2$ is sent by this inclusion to the required fundamental group of the complement to the hypersurface Σ in $\mathbb{C}^n$.

Zariski Theorem 2. *The map induced by the above inclusion,*

$$\pi_1(\mathbb{C}^2 \smallsetminus \Sigma, *) \longrightarrow \pi_1(\mathbb{C}^n \smallsetminus \Sigma, *)\,,$$

is an isomorphism. In other words, all relations in the fundamental group $\pi_1(\mathbb{C}^n \smallsetminus \Sigma, *)$ are realisable by homotopies of loops lying in the fibre $\mathbb{C}_*$ which avoid Σ and do not leave $\mathbb{C}^2$.

To understand it, project everything to $\mathbb{C}^{n-1}$ (along $\mathbb{C}_*$). The projection of the pair $(\Sigma \subset \mathbb{C}^n)$ would be fibred over $\mathbb{C}^{n-1}$ if we delete from the base $\mathbb{C}^{n-1}$ the *discriminant hypersurface* (formed by the nongeneric fibres).

Applying Zariski Theorem 1 to this discriminant hypersurface of $\mathbb{C}^{n-1}$, we can homotopically transform any loop of its complement in $\mathbb{C}^{n-1}$ into a loop of its complement inside a generic complex line of $\mathbb{C}^{n-1}$.

The preimage of this complex line of $\mathbb{C}^{n-1}$ is a complex 2-plane of $\mathbb{C}^n$. Zariski Theorem 2 provides a homotopic deformation of any relation in the required fundamental group to the part of $\mathbb{C}^n \setminus \Sigma$ in this complex 2-plane.

Therefore, Zariski Theorem 2 reduces the study of the fundamental group of the complement to an algebraic hypersurface Σ of $\mathbb{C}^n$ to the study of the fundamental group of the complement to the algebraic curve $\widehat{\Sigma} = \Sigma \cap \mathbb{C}^2$ of $\mathbb{C}^2$ (whose generators are provided by Zariski Theorem 1).

To complete the picture, there is a third theorem of Zariski (stated below), which provides an explicit description of the relations between the generators (provided by Theorem 1) of the fundamental group of the complement to an algebraic curve $\widehat{\Sigma}$ in the plane $\mathbb{C}^2$.

We shall denote the complex coordinates in $\mathbb{C}^2$ by (x, y), so that the projection p sends the point (x, y) to x.

Recall that the generators $\gamma_1, \dots, \gamma_\mu$ of the group $\pi_1(\mathbb{C}^2 \setminus \Sigma, *)$ are the simple loops going around the respective points $\alpha_1, \dots, \alpha_\mu$ of $\Sigma \cap \mathbb{C}_*$ – Fig. 5.23.

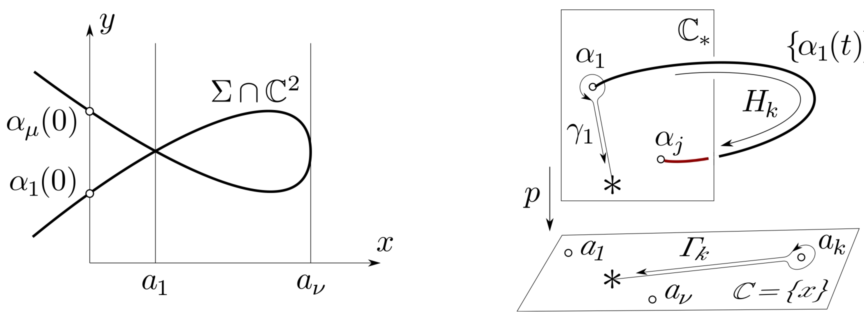


Figure 5.23: Action H_k of the loop Γ_k (of the space of generic x-points) on the basic fibre $\mathbb{C}_*$ and on its intersection points α_ℓ with curve Σ.

Denote by $\{a_1, \dots, a_\nu\}$ the projections of the exceptional fibres. The fundamental group F_ν of the complement to the points $\{a_1, \dots, a_\nu\}$ in the base line $\{x\} = \mathbb{C}$, is generated by the simple loops $\Gamma_1, \dots, \Gamma_\nu$ which go around the corresponding points $\{a_1, \dots, a_\nu\}$:

$$\pi_1\big(\mathbb{C} \smallsetminus \{a_k\}, p(*)\big) \approx F_\nu \ .$$

Over the points of the loop $x = \Gamma_k(t)$ every fibre contains μ points $\alpha_1(t), \dots, \alpha_\mu(t)$ of Σ continuously varying with t. Hence for each loop $x = \Gamma_k(t)$ we define a continuous family of homeomorphisms $h_k(t)$, sending the fibre $\mathbb{C}_*$ together with its points $(\alpha_1, \dots, \alpha_\mu)$ to the moving fibre over $x = \Gamma_k(t)$ with its respective moving points $(\alpha_1(t), \dots, \alpha_\mu(t))$. At the end we get the homeomorphism $H_k = h_k(1)$ (defined by the loop Γ_k).

We may choose our base point $*$ in such a way that H_k (and even h_k) be the identity map in a neighbourhood of this point (being the identity map in the whole domain $|y| \geqslant |y(*)|$).

Applying the homeomorphism H_k to the loops $\gamma_1, \dots, \gamma_\mu$, we obtain μ new elements $H_{k*}\gamma_1 \ \dots, \ H_{k*}\gamma_\mu$ of the free group F_μ. This map $H_{k*} : F_\mu \longrightarrow F_\mu$ is an isomorphism of the fundamental group of the complement $\mathbb{C}_* \smallsetminus \Sigma$ to itself, induced by the homeomorphism H_k (for each $k \in \{1, \dots, \nu\}$).

Thus, in the required fundamental group $\pi_1\big(\mathbb{C}^2 \smallsetminus \Sigma, *\big)$ the generators γ_ℓ verify the relation

$$R_{k,\ell} \ : \quad (H_{k*}\gamma_\ell)\,\gamma_\ell^{-1} = 1\,. \tag{6}$$

Of course, $H_{k*}\gamma_\ell$ is a word in terms of the generators $\gamma_1, \dots, \gamma_\mu$, and hence $R_{k,\ell}$ is indeed a relation.

Zariski Theorem 3. *The fundamental group of the complement to a complex algebraic curve Σ in the complex plane $\mathbb{C}^2$ is the group generated by the generators $\gamma_1, \dots, \gamma_\mu$, with the defining relations $R_{k,\ell}$ $(1 \leqslant k \leqslant \nu,\ 1 \leqslant \ell \leqslant \mu)$.*

Remark. Algebraically this geometric fact means the natural isomorphism

$$\pi_1(\mathbb{C}^2 \smallsetminus \Sigma, *\big) \approx \frac{F_\mu}{R} \ ,$$

where R is the normal subgroup of F_μ "defined by relations (6)", that is, R is the smallest subgroup of F_μ containing the elements $(H_{k*}\gamma_\ell)\,\gamma_\ell^{-1}$ together with all their conjugates $f(H_{k*}\gamma_\ell)\,\gamma_\ell^{-1}f^{-1}$, where $f \in F_\mu$ (see p. 53).

The fulfilment of the relations $R_{k,\ell}$ is proved above. The remaining part of Theorem 3 –stating that they form a complete system of relations: *All other relations are their algebraic corollaries*– is explained in this remark.

Example (Complement of the Semicubic Parabola). We shall apply Zariski Theorem 3 to compute the fundamental group of the complement to the semicubic cusp $\Sigma = \{y^2 = x^3\}$. We choose the basic fibre $\mathbb{C}_* : x = 1$.

In this example, there are two points α_ℓ ($y = \pm 1$), one point a_k (at $x = 0$) and one loop Γ_k (which can be chosen to be $x = e^{2\pi i t}$) . The two moving points $\alpha_\ell(t)$ have the y-coordinates

$$y_\ell = \pm e^{(3/2 \cdot 2\pi i t)} .$$

Thus, it is easy to construct explicitly the homeomorphisms h_t which move these two points in the prescribed way. These homeomorphisms are shown in Fig. 5.24 (for the 6 values $t = 1/6, 2/6, \ldots, 1$).

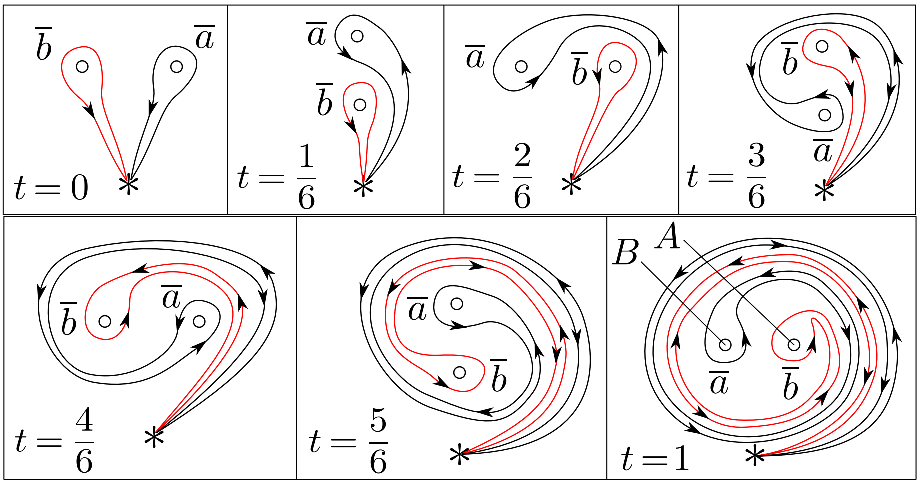


Figure 5.24: The family of the homeomorphisms $h_t : \mathbb{C} \to \mathbb{C}$, moving the two points $y = \pm 1$ to the two points where $y = \pm \exp(2\pi i t)3/2$, and these homeomorphisms action on the loops a and b, transformed to $\bar{a}$ and to $\bar{b}$.

The loops a and b are transformed by the homeomorphisms h_t to new loops, $\bar{a} = h_t a$ and $\bar{b} = h_t b$. For $t = 1$, when the pair of points $y_\ell(t)$ returns to its initial place, the transformed loops $\bar{a}$ and $\bar{b}$ define two elements of the fundamental group F_2 of the complement to these points. To represent these elements of the fundamental group as words in terms of the generators a and b, we use the standard method (explained on p. 120, Fig. 4.7) of counting their intersections with the lines that connect the excluded points to infinity.

In this way we get from Fig. 5.24 the expressions

$$\bar{a} = h_1 a = ababa^{-1}b^{-1}a^{-1} , \quad \bar{b} = h_1 b = abab^{-1}a^{-1} .$$

Consequently, the relations (provided by Zariski Theorem 3) on the fundamental group of the complement to the semicubic cusp in the complex plane are

$$ababa^{-1}b^{-1}a^{-1} = a\ , \quad abab^{-1}a^{-1} = b\ .$$

Both relations mean the same homotopy, $aba = bab$, $bab = aba$, which is just the standard relation of the 3-string braid group (see also p. 139):

$$\mathrm{Br}(3) \approx \pi_1\big(\mathbb{C}^2 \smallsetminus \{y^2 = x^3\}, *\big)\ .$$

The generators a and b with their relation $aba = bab$ can be replaced by the generators $p = ab$ and $q = aba$ with their relation $p^3 = q^2$ (since $(ab)(ab)(ab) = (aba)(aba)$ if and only if $aba = bab$).

This "strange algebraic coincidence" reflects important geometrical relations of braid theory with elliptic curves (as well as with their higher genus generalisations).

5.9 Monodromy representation $\mathrm{Br}(3) \to \mathrm{SL}(2, \mathbb{Z})$

Consider an element (z_1, z_2, z_3) of the configuration space $\mathrm{Conf}_3(\mathbb{R}^2 \approx \mathbb{C})$ of three unordered points of the plane. We associate to this configuration element the elliptic curve

$$y^2 = (x - z_1)(x - z_2)(x - z_3)$$

in $\mathbb{C}\mathrm{P}^2$ (with x and y being its affine coordinates). This curve is a two-fold ramified covering of the sphere $\mathbb{S}^2 = \mathbb{C}\mathrm{P}^1$ with affine coordinate x. It is ramified at infinity and at the three points of our configuration.

Monodromy. A loop γ in the space $\mathrm{Conf}_3(\mathbb{C})$ generates a continuous one-parameter family of elliptic curves, which starts and ends at the same elliptic curve. Identifying homeomorphically (by small homeomorphisms) the neighbouring tori forming this family, we get a homeomorphism of the initial torus to itself, called "*monodromy* of the family".

At least up to homotopy, this monodromy homeomorphism h_γ can be represented by an element $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ of the modular group $\mathrm{SL}(2, \mathbb{Z})$; namely, it is homotopic to the map of the torus onto itself which is defined, in terms of the angular coordinates on $\mathbb{S}^1 \times \mathbb{S}^1$, by the formula

$$h_\gamma \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} a\alpha + b\beta \\ c\alpha + d\beta \end{pmatrix}\ .$$

To see this, observe that the monodromy representation under consideration, $h : \mathrm{Br}(3) \to \mathrm{SL}(2, \mathbb{Z})$, is the action of the braid group on $\pi_1(\mathbb{T}^2) = \mathbb{Z}^2$.

Now, if we glue the torus from a square, then the sides become loops representing two generators of $\pi_1(\mathbb{T}^2)$. Identifying this fundamental group with $\mathbb{Z}^2$, we represent these generators by $(1, 0)$ and $(0, 1)$, respectively. The images (a, c) and (b, d) of these generators under the monodromy homeomorphism form the columns of the monodromy matrix*.

Thus we get a homomorphic *monodromy representation* h of the 3-string braid group,

$$h : \mathrm{Br}(3) \longrightarrow \mathrm{SL}(2, \mathbb{Z}), \tag{7}$$

in which the braid defined by the loop γ is represented by the element h_γ of the modular group.

Proposition. *The monodromy representation* (7) *is surjective and its kernel is non-trivial.*

This fact provides an additional relation in the modular group, missing in the braid group for the generators a, b (or p, q defined on p. 161).

To prove it, we shall calculate the two matrices h_a and h_b which represent the generators a and b of the braid group $\mathrm{Br}(3)$ in the matrix group $\mathrm{SL}(2, \mathbb{Z})$.

To calculate these matrices, we shall use the basic closed curves on the torus surface of the elliptic curve ("the parallel and the meridian"), representing them on the base plane of complex numbers $\mathbb{C} = \{x\}$, covered two-fold by the torus with ramifications at z_1, z_2, z_3 and at infinity – Fig. 5.25.

Namely, we cut the x-plane by two disjoint segments – one joins z_1 to z_2, the other joins z_3 to ∞. This operation represents the torus surface as two copies of the x-plane, with two cut segments on each of them – the "first sheet" and the "second sheet" of the surface. The whole torus is represented as the union of these two sheets, glued along the cuts in such a way that each side of the cut on the first sheet is glued with its opposite side on the second (a moving point crossing the cut of the surface, changes the sheet).

The resulting surface is everywhere smooth and oriented (by the growth direction of the argument of the complex numbers). The unusual behaviour near the ramification points is due to the fact that x is not a regular coordinate at these places; one should use the y coordinate to see the smoothness.

---

*It is known that a homeomorphism on $\mathbb{T}^2$ is determined uniquely, up to homotopy, by its action on $\pi_1(\mathbb{T}^2)$ (but we will not prove it in this book). A similar assertion holds for a surface of any genus (see the next section).

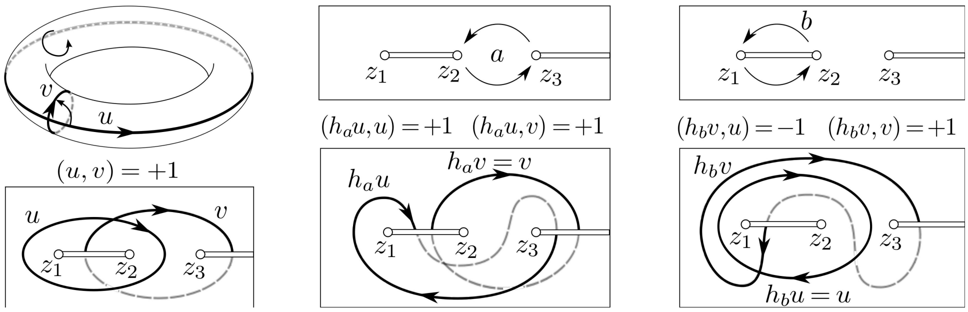


Figure 5.25: Monodromy actions h_a and h_b of the braid group generators a and b on the basic curves u and v of the torical surface of the elliptic curve, ramified over the x-plane at (z_1, z_2, z_3, ∞).

The basic curves u and v are represented in the x-plane by two closed curves: u is the boundary of a disc connecting the points z_1 and z_2 (and containing the joining cut segment), while v is the boundary of a similar disc joining the point z_2 to the point z_3.

Thus, the curve u remains on the sheet of the torical surface where it had started (never crossing the cuts), let it be the first sheet.

The other curve, v, cross both cuts and is subdivided into two parts, one on the first sheet, the other on the second one. The part on the second sheet is represented in Fig. 5.25 by a broken line. The curve v intersects the first curve once, at the first sheet (the second intersection point on the x-plane —visible in Fig. 5.25— does not represent a real intersection of the curves on the covering torus, but it is only an intersection of their projections).

Every closed curve ξ on the torus is homotopic to an integral linear combination of the curves u and v: $\xi \sim mu + nv$, $m \in \mathbb{Z}$, $n \in \mathbb{Z}$.

To find the coefficients m and n, the best way is to calculate the "intersection numbers" (ξ, u) and (ξ, v).

Intersection Index. The *intersection index* (ξ, η) of two closed oriented curves ξ and η on an oriented surface, is defined as the sum of $+1$ s and -1 s associated to their intersection points in the following way.

We suppose that the curves are smooth and intersect transversely (with no tangency). An intersection point contributes $+1$, if the pair of velocity vectors of the first and of the second curve, at that point, orient the surface positively. The contribution is -1 in the opposite case.

The sum of these contributions remains constant under a homotopy of the curves, and thus the integer (ξ, η) depends only on the homotopy classes of the curves.

For example, the intersection index of our curves u and v is equal to $+1$, while the self-intersection numbers (u,u) and (v,v) vanish. This follows, for instance, from the evident anti-symmetry $(\xi,\eta) = -(\eta,\xi)$.

Thus, the intersection index of a curve $\xi \sim mu + nv$ with v is equal to m and with u is equal to $-n$. Hence we can write $\xi \sim (\xi,v)\,u - (\xi,u)\,v$.

Knowing the homeomorphism h_γ, we draw the image $\overline{w} = h_\gamma w$ of a closed curve w, projected to the x-plane. Intersecting the curve w on the torus with the initial curves u and v, we get the representation of the resulting homotopy class

$$w \sim (\overline{w},v)\ u - (\overline{w},u)\ v\ .$$

Applying this construction to the loops γ, defining the generators a and b of the 3-braid group, and to the initial curves $w = u$ and $w = v$, we get from Fig. 5.25 (for the standard basis (u,v) of the curves on the torus surface):

$$h_a u = u - v\ , \qquad h_a v = v\ ,$$
$$h_b u = u\ , \qquad h_b v = u + v\ ,$$

from which we get the two monodromy matrices,

$$h_a = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}\ , \qquad h_b = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\ .$$

This proves that the monodromy representation h covers the modular group[*].

Therefore the monodromy matrices of the generators p and q are

$$h_p = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\ , \qquad h_q = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\ .$$

They verify, of course, the braid group relation $h_p^3 = h_q^2$ and we have an additional relation in $\mathrm{SL}(2,\mathbb{Z})$: $h_q^2 = -1$.

---

[*]The fact that the two triangular matrices h_a and h_b generate the whole modular group $\mathrm{SL}(2,\mathbb{Z})$ has a simple geometrical meaning, describing a way to construct any unit area integer parallelogram in the lattice $\mathbb{Z}^2$ from the basic square.

Consider the following operation on the parallelogram formed by the vectors ξ and η: one of these vectors remains, but we modify the other vector by adding to it the first one (or we subtract it for the opposite operation). The new vectors (say, ξ and $\eta + \xi$) define a new integer parallelogram of area 1.

The geometric fact we are discussing is that each area 1 parallelogram with integer vertices is attainable from the basic square of the lattice by a finite chain of such operations (preserving sometimes the first vector of the pair, sometimes the second).

5.10 Fundamental groups of the closed surfaces

Problem. Compute the fundamental group of the surface of genus g (of a sphere with g handles).

Answer. This group has a system of $2g$ generators (a_k, b_k) with the defining relation

$$a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} \ldots a_g b_g a_g^{-1} b_g^{-1} = 1 \ . \tag{8}$$

Solution. To obtain the generators, one cut the surface by making two circular cuts on each handle, one along the "parallel" a_k and one along the "meridian" b_k, like one cut the torus to obtain a rectangle from which the torus can be reconstructed by gluing the corresponding sides generated by the same circular cutting.

For $g = 2$ this cutting is shown in Fig. 5.26. The resulting $4g$-gone suffices to reconstruct the initial surface by gluing the corresponding sides.

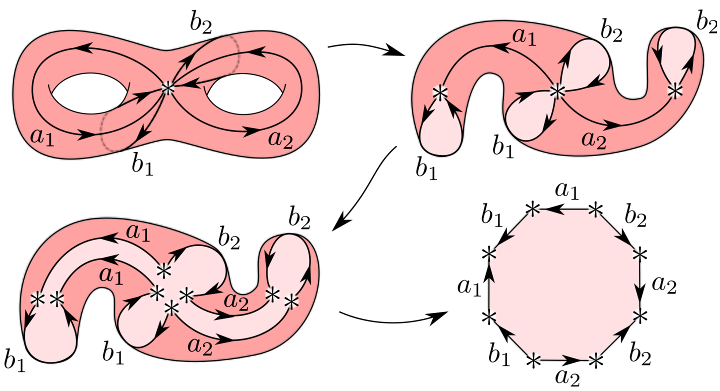


Figure 5.26: The cutting of a surface of genus $g = 2$, making from it a $4g$-gone.

These $4g$ oriented sides represent $2g$ loops on the initial surface, and the $4g$ vertices of the $4g$-gone are all glued to the basic point $*$ of our fundamental group (it is then the initial and final point of the above $2g$ loops).

Proposition. *The homotopy class of any loop γ is representable as a product of the classes of the edges, a_k, b_k, a_k^{-1}, b_k^{-1}.*
(So, these classes are generators of $\pi_1(M_g)$.)

Proof. Any loop γ has a homotopic representative, which is a smooth map of the segment $[0, 1]$ to our surface, starting and ending at $*$. It can intersect the simple curves a_k and b_k many times, but we can find a homotopic representative for which there will be only a finite number of simple crossings (with no tangencies). Moreover, we can deform the loop γ (inside its homotopy class) in such a way that every crossing occurs at the basic point $*$.

These crossing moments subdivide the loop into a finite sequence of segments that are sent to the surface in such a way that both boundary points are sent to the basic point $*$, while the interior part of those segments is sent to the uncut part. Such segment is a curve in the $4g$-gone, which starts and ends at two vertices and has no other common point with the boundary, being then homotopic to the sequence of the edges joining the same vertices. □

Proposition. *Any relation in the fundamental group of a surface of genus g is generated by relation* (8).

Although a formal proof of this geometric statement would be a good exercise, we present the proof here.

Proof. The homotopy of the product (8) to the trivial loop follows from the contractibility of the boundary of the $4g$-gone (to any point inside it).

A defining relation in $\pi_1(M_g)$ means that two loops, generated in some given way, are homotopic, that is, there is a homotopy between them. Now, deforming that homotopy to a smooth one, $F : I^2 \to M_g^2$, one consider the preimage $F^{-1}(\partial)$ of the union ∂ of our cuts a_k and b_k in the square I^2.

This preimage subdivides the square I^2 into pieces, say ℓ pieces, whose boundaries are sent by F to ∂, and whose interior parts do not meet ∂.

The decomposition of I^2 into these ℓ pieces leads to the replacement of F by an equivalent product of ℓ "local" homotopies, $F_j : I_j^2 \to M_g$, each of which sends the interior part of I^2 to the complement to the cut set ∂. Then each of these elementary homotopies defines a homotopy of a path along the cut set ∂ to the interior part $M_g^2 \setminus \partial$ of our $4g$-gone.

Such an elementary homotopy is reducible to relation (8) by the elementary geometry of the interior of the $4g$-gone, diffeomorphic to the standard disc or square. Returning to the initial homotopy F, we see that it is reducible to a finite product of the elementary ones, each of which is reducible to relation (8), which generates, therefore, the homotopy F. □

5.11 Fibred Surfaces and Fibred Manifolds

Problem. Find the surfaces fibred non-trivially into circles, such that the base B of the fibration is also a circle. Calculate their fundamental group.

Solution. The only such surface (up to diffeomorphisms in the smooth case and up to homeomorphisms in the C^0-case) is the *Klein bottle*.

Any fibration whose base is a circle can be described as a fibration over the segment $I = [0, 1]$, whose end fibres, F_0 and F_1, are identified ("glued") by some diffeomorphism (or by a homeomorphism in the C^0-case).

Different identifications produce many manifolds, but they are diffeomorphic when the corresponding identifications are homotopic.

For the homeomorphisms (or diffeomorphisms) of the circle there are two homotopy classes: all homeomorphisms that preserve the orientation belong to the homotopy class of the identity map; all the orientation reversing homeomorphisms belong to the class of the reflection of the circle.

Thus there are exactly two non-diffeomorphic fibrations over the circle whose fibres are circles: gluing the end-fibres $\mathbb{S}^1_0$ and $\mathbb{S}^1_1$ by the identity map we get the direct product $T^2 = \mathbb{S}^1 \times \mathbb{S}^1$ (the torus); while the gluing by the reflection of the circle provides a different surface - the *Klein bottle*.

The difference between the Klein bottle and the torus is very similar to the difference between the cylinder (which is the direct product $\mathbb{S}^1 \times I$) and the Möbius band (which is also fibred over the circle $\mathbb{S}^1$ with fibres I, but whose gluing of the initial and final fibres, F_0 and F_1, reverses the orientation of the factor I of the product).

Thus, like the torus, one can represent the Klein bottle by using the square with identified opposite sides. The difference is shown in Fig. 5.27, where the two horizontal sides are glued with reversed orientation.

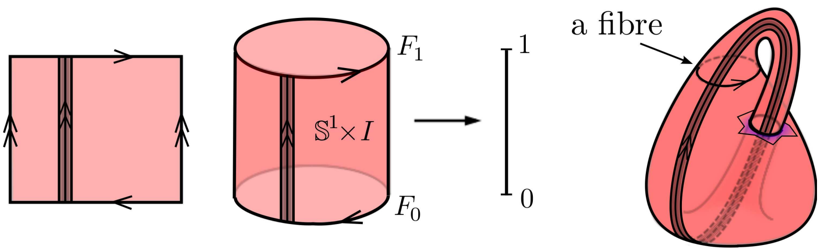


Figure 5.27: Klein bottle's construction, and a smooth immersion of it in $\mathbb{R}^3$.

The resulting smooth surface is nonorientable because it includes the Möbius band, formed by (small) arcs, one at each fibre-circle. Unlike the Möbius band, the Klein bottle cannot be embedded into Euclidean 3-space. However, it can be immersed (locally embedded, allowing self-intersection – in the same way as the curve "∞" represents an immersion of a circle into the plane) – Fig. 5.27. The immersed Klein bottle is not good as a bottle because it does not separate the "interior" part from the outer space.

The fundamental group of the Klein bottle can be described as the group of pairs of integers (m, n) with the following strange multiplication rule:

$$(m_1, n_1) \cdot (m_2, n_2) = (m_1 + m_2, n_1 + (-1)^{m_1} n_2) \ .$$

Problem. Calculate the fundamental group of the 3-manifolds fibred into tori, and whose base manifold is the circle $\mathbb{S}^1$.

Hint. Suppose the Manifold M and its fibration into tori are obtained from the product $[0,1] \times \mathbb{T}^2$, by gluing the fibres over the end-points of the segment $[0,1]$, and suppose that the gluing is described by a given matrix $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ of $\mathrm{SL}(2,\mathbb{Z})$:

$$(u, v) \mapsto (au + bv, cu + dv),$$

for the angular coordinates u, v on $\mathbb{T}^2$. Denote by M_A this 3-manifold.

The problem is to relate the properties of the fundamental group $\pi_1(M_A)$ to those of the matrix A describing the gluing.

Problem. Let M_g^2 be a surface of genus g equipped with some Riemannian metric*. Calculate the fundamental group of the 3-manifold $T_1 M_g^2$ of all tangent vectors of length one to the surface M_g^2.

Answer. The fundamental group $\pi_1(T_1 M_g^2)$ has the system of $2g+1$ generators $(a_1, b_1, \ldots, a_g, b_g, c)$ with the defining relation

$$a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} \ldots a_g b_g a_g^{-1} b_g^{-1} \ = \ c^{2-2g} \,, \tag{9}$$

where c is a new generator that commutes with the standard generators $(a_1, b_1, \ldots, a_g, b_g)$ of $\pi_1(M^2)$.

---

*A *Riemannian metric* on a differentiable manifold M is a positive-definite symmetric bilinear form (*inner product*) $\langle \cdot, \cdot \rangle$ on every tangent space $T_x M$. A Riemannian metric enables one to measure the angle between two tangent vectors at a given point of M, and the lengths of the vectors: $|\xi| = \sqrt{\langle \xi, \xi \rangle}$.

For example, the length of a curve on a manifold, $\gamma : [t_0, t_1] \to M$, is expressed using the metric as $\ell(\gamma) = \int_{t_0}^{t_1} \sqrt{\langle \dot\gamma, \dot\gamma \rangle} dt$.

Hint. The generator c represents the fibre circle. The other $2g$ generators are vector fields along the standard $2g$ loops generating $\pi_1(M_g^2)$.

For most compact surfaces M_g^2 is not possible to continue these fields inside the $4g$-gone without zeros. Relation (9) expresses, in terms of the genus, the "mesure" of the obstruction to such a continuation.

Namely, given a vector field nowhere vanishing outside a small disc of M_g^2 the obstruction to continue the field without zeros inside the disc is "measured" by counting the number of turns of the field of directions when one makes one positive tour along the boundary of the disc.

This number is the Euler characteristic $\chi(M_g) = 2-2g$. This follows from the fact that the sum of indices of the singular points of a smooth vector field on a surface of genus g is equal to $2 - 2g$ (Ch. 10, pp. 372-374).

Thus for $g = 1$ there is no obstruction (because $\mathbb{T}^2$ is parallelisable) :

$$T_1(\mathbb{T}^2) = \mathbb{T}^3 \quad \text{and} \quad \pi_1(T_1(\mathbb{T}^2)) = \mathbb{Z}^3 .$$

For a surface of genus $g \neq 1$ the "level of obstruction" to the continuation is the Euler characteristic $\chi = 2 - 2g$.

In the case of the sphere, $g = 0$, such a field of directions should make two positive turns when one makes one tour in the positive direction along the boundary of the disc (see Fig. 10.9 on p. 372).

The existence of such a field follows from the $4g$-gone description of the surface (one may use the "constant" vector field, invariant by translations, of the plane containing the $4g$-gone).

Problem. Calculate the fundamental group of the space of ordered configurations of three points of $\mathbb{C}$,

$$\pi_1\bigl(\mathbb{C}^3 \smallsetminus \{(x-y)(y-z)(z-x) = 0\}, *\bigr) .$$

Colored Braids. The fundamental group of the space of ordered configurations of n points of $\mathbb{C}$ is called the *colored braid group* of n strings. Its loops are particular braids distinguished from the general ones by the fact that each string of the braid ends at the same place where it started.

Equivalently, the colored braids are defined as the braids that determine the trivial permutation of the points of the initial configuration.

For $n = 3$ this colored braid group is closely related to the modular group $\mathrm{SL}(2, \mathbb{Z})$.

5.12 On the Zariski "Braided Syzygies"

Discriminant The Zariski description of the fundamental group of an algebraic hypersurface Σ in the complex space $\mathbb{C}^n$ in terms of the projection $p : \mathbb{C}^n \to \mathbb{C}^{n-1}$ provides an algebraic hypersurface Δ (of one dimension less than Σ) in the base space $B = \mathbb{C}^{n-1}$ of the projection. This hypersurface is called the *discriminant* of the restriction $p_{|\Sigma} : \Sigma \to \mathbb{C}^{n-1}$ of the projection p to the hypersurface Σ: The union of the fibres $F_b = p^{-1}(b)$ which are non-generic (with respect to Σ) is projected by p to this algebraic hypersurface $\Delta \subseteq B$.

Over the complement to the discriminant hypersurface the projection $p|_\Sigma$ is a covering: for each $b \in B \setminus \Delta$ the number μ of intersection points of the fibre F_b with Σ is the degree of the hypersurface Σ.

We can iterate the construction, studying the fundamental group of the complement to the hypersurface Δ in B by projecting $B = \mathbb{C}^{n-1}$ generically to $\mathbb{C}^{n-2}$ and taking the discriminant of the restriction of this projection to the hypersurface Δ, and so on.

This iteration provides a series of projections $p_j : E_j \to B_j$, $j = 1, 2, \ldots$ of complex spaces $E_j = \mathbb{C}^{n+1-j}$ to smaller complex spaces $B_j = \mathbb{C}^{n-j}$, with 1-dimensional fibres $F_b^j = p_j^{-1}(b)$ $(b \in B_j)$. Each space E_j contains an algebraic hypersurface Σ_j and the object to study is the fundamental group $G_j = \pi_1(E_j \setminus \Sigma_j)$ of its complement.

Applying the 3rd Zariski theorem to its study, we consider the following objects.

The generic fibre $F_*^j \approx \mathbb{C}$ intersects the hypersurface Σ_j at μ_j points $(\alpha_1^j, \ldots, \alpha_{\mu_j}^j)$. The fundamental group of their complement in the fibre

$$\pi_1\big(F_*^j \setminus \{\alpha_1^j, \ldots, \alpha_{\mu_j}^j\}, *\big)$$

is the free group $\mathcal{F}_{\mu_j}$ with generators $(\gamma_1^j, \ldots, \gamma_{\mu_j}^j)$. Here the simple loop of the class γ_k^j comes from the base point $*$ to the intersection point α_k^j, turns once around this intersection point and returns to the base point along the same joining path.

The homotopy classes of the "vertical loops" γ_k^j generate the whole fundamental group G_j (by the first Zariski theorem).

The loops γ_k^{j+1} generate the (lower) fundamental group G_{j+1} of the complement to the discriminant $\Delta_j = \Sigma_{j+1}$ in the base space $B_j = E_{j+1}$, while for the preceding (upper) fundamental group G_j they generate its relations, according to the third Zariski theorem.

Namely, the braided monodromy M_j associates to each element γ of the lower fundamental group $G_{j+1} = \pi_1\big(B_j \smallsetminus \Delta_j, *\big)$ –for instance, to any generator $\gamma_\ell^{j+1} \in \mathcal{F}_{\mu_{j+1}}$– the braid $M_j[\gamma] \in \mathrm{Br}(\mu_j)$, defined by the covering motion of the configuration $\{\alpha_k^j\}$ over the motion γ_ℓ of the base point.

This μ_j-string braid acts on the complement to the points $\{\alpha_k^j\}$ in the fibre $F_*^j \approx \mathbb{C}$, and induces the monodromy automorphism $T_\gamma : \mathcal{F}_{\mu_j} \to \mathcal{F}_{\mu_j}$.

The relation in the group G_j (associated to the element γ of the fundamental group G_{j+1} of the base) is that, in the group G_j, $T_\gamma\varphi \sim \varphi$ for any $\varphi \in \mathcal{F}_{\mu_j}$.

It means that the element

$$r_{k,\ell}^j = \big(T_{\gamma_\ell^{j+1}}(\gamma_k^j)\big)(\gamma_k^j)^{-1}$$

of the free group $\mathcal{F}_{\mu_j}$ belongs to the normal subgroup $R_j \subset \mathcal{F}_{\mu_j}$ of the relations of the upper fundamental group $G_j = \mathcal{F}_{\mu_j}/R_j$.

The normal subgroup R_j is the minimal normal subgroup of the group $\mathcal{F}_{\mu_j}$, containing all the $\mu_j\mu_{j+1}$ elements $r_{k,\ell}^j$.

In this sense *the generators γ_ℓ^{j+1} of the next group G_{j+1} define the relations $r_{k,\ell}^j$ of the group G_j.*

Similarly, the *relations* of the next group, G_{j+1}, are produced from the generators of G_{j+2}, defining by the braided monodromy the elements $r_{k,\ell}^{j+1}$ of the free group $\mathcal{F}_{\mu_{j+1}}$.

In terms of the group G_j these "relations between its relations" form some kind of "syzygies". Namely, since the element $r_{k,\ell}^{j+1}$ is a monomial in the generators γ_m^{j+1} of the free group $\mathcal{F}_{\mu_{j+1}}$, it has a braided monodromy action

$$T_\gamma = \prod_k T^{\pm 1}_{\gamma_{m_k}^{j+1}},$$

which is an automorphism of $\mathcal{F}_{\mu_{j+1}}$ to itself.

This automorphism is "a braided relation between the relations of the group G_j", since it transforms every relation $r \in R_j$ to itself.

It means that replacing in the monomial

$$\varrho_{\gamma,\gamma_\ell} = \big(T_\gamma(\gamma_\ell^{j+1})\big)(\gamma_\ell^{j+1})^{-1} \in \mathcal{F}_{\mu_{j+1}}$$

(expressed in terms of the generators of the free group $\mathcal{F}_{\mu_{j+1}}$) each generator γ_m^{j+1} by the corresponding element $r_{k,m} \in \mathcal{F}_{\mu_j}$, we would obtain the trivial element 1 of the free group $\mathcal{F}_{\mu_j}$.

This description suggests the following new question. Interpreting the relations of the next group, G_{j+1}, as some braided relations between the relations of the preceding group G_j, it would be interesting to study the whole group of such syzygies. It contains the generators of the next fundamental group, but there could exist other independent syzygies. Thus the question is to study the quotient group of the group of all the braided syzygies of the group G_j and the group of relations R_{j+1} of the group G_{j+1} producing the relations of G_j.

Even in the case of the braid group $\mathrm{Br}(n)$ (where the initial hypersurface is the $n-1$ dimensional swallowtail in $\mathbb{C}^n$) we have that, for a generic sequence of the projections $p_j : E_j \to B_j$, these groups of syzygies are unknown. Hence it remains some hope that in this particular case all the braided syzygies of the groups G_j can be reducible to the subgroup R_{j+1} of the relations of the group G_{j+1}, producing the relations of G_j (by the 3rd Zariski theorem).

Chapter 6

Integration

To understand the geometry of the integration on manifolds, one has first to recognise that the "integration of functions" in the courses of calculus is a nonsense.

Indeed, even the Leibniz notation

$$\int f(x)\,dx$$

suggests it: the object to integrate is rather the expression $f(x)\,dx$, than "the function f".

One should never forget that the "integral of the function f" depends also on the choice of the coordinate x on the domain of definition. If we made a diffeomorphism (say, $\widetilde{x} = x^2$ on the positive semi-axis) we would then introduce a new coordinate which would change completely the integral of the same function, having the prescribed values at the points of the domain of definition, while the integrals

$$\int f(x)\ dx\ , \qquad \int f\left(\sqrt{\widetilde{x}}\right) d\widetilde{x}\ ,$$

along the same geometric domain, take different values.

The situation is similar, but even more complicated, in the case of higher-dimensional domains (needed for the double integrals, triple integrals, vector-analysis and so on).

To start the theory of integration on manifolds we have to eliminate the coordinate system from the initial definition. To do it, we need first some elementary algebra (missing, unfortunately, in the present day elementary courses at the universities).

6.1 Exterior forms

Consider a real vector space $V \approx \mathbb{R}^n$ with the usual linear coordinate system $\{x_k : V \to \mathbb{R},\ 1 \leqslant k \leqslant n\}$.

Definition. A *form* (or 1-*form*) on V is a linear map from V to $\mathbb{R}$.

Theorem 1. *The space of* 1*-forms is a vector space of the same dimension* n *as* V (*for the ordinary operations of the addition of forms and of their multiplication by numbers*).

Proof. We shall prove that the n 1-forms provided by the coordinates x_k, do form a basis of the space of the 1-forms on V.

Consider a 1-form $\alpha : V \to \mathbb{R}$ and write $e_1, \ldots, e_n$ for the basis of V defining the coordinate system $x_1, \ldots, x_n$. Denote by a_k the value of the form α on the vector $e_k \in V$, and then construct the form

$$\omega = a_1 x_1 + \cdots + a_n x_n .$$

The values of ω on the basic vectors are $\omega(e_k) = a_k$, and hence (by linearity) both forms α and ω are equal everywhere. Thus, every 1-form is a linear combination of the n 1-forms x_k.

These n 1-forms are linearly independent, since for any linear combination of them which vanishes identically,

$$b_1 x_1 + \cdots + b_n x_n \equiv 0 \ ,$$

its coefficients b_k (which are its values on the basic vectors e_k) vanish, $b_1 = \cdots = b_n = 0$. Consequently, the n 1-forms $x_1, \ldots, x_n$ are linearly independent and thus form a basis of the vector space of the 1-forms on V. □

The vector space of the 1-forms on V is called *the dual space* of V and is denoted by V^* (or $\mathbb{R}^{n*}$ for $V = \mathbb{R}^n$).

The use of the word “dual” here is justified by the important (but easy)

Theorem 2. *The dual space of the dual to some original finite dimensional vector space is the original vector space:*

$$V^{**} = V \ .$$

The theorem follows from the fact that the value $a_1x_1 + \cdots + a_nx_n$ of the form a (with coefficients a_k) on the vector x (with components x_k) is linear in both arguments a and x, and that the map $V \times V^* \to \mathbb{R}$, $(x, a) \mapsto a_1x_1 + \ldots + a_nx_n$, depends on both arguments x and a symmetrically.

We shall need also some special functions of two or more vectors of the vector space V.

Definition. A function ω of k vectors is called a *k-form* (or an *exterior k-form*), if it verifies the following two conditions:

1) it is linear* with respect to any of the k vectors, defining a k-linear map

$$\omega : \underbrace{V \times V \times \cdots \times V}_{k} \longrightarrow \mathbb{R} \ .$$

2) it is anti-symmetric, changing its sign under any odd permutation of the arguments and preserving it under any even permutation.

Example. For $k = 3$ the antisymmetry condition is

$$\omega(x, y, z) = \omega(y, z, x) = \omega(z, x, y) = -\omega(y, x, z) = -\omega(x, z, y) = -\omega(z, y, x)$$

for any three vectors x, y, z.

Of course, the anti-symmetry under the transpositions of two elements is sufficient for the general antisymmetry property

$$\omega(\xi_{i_1}, \ldots, \xi_{i_k}) = (-1)^{s(i_1,\ldots,i_k)} \ \omega(\xi_1, \ldots, \xi_k) \ ,$$

where the "parity" $s(i_1, \ldots, i_k)$ of the permutation $(i_1, \ldots, i_k)$ of k elements $(1, \ldots, k)$, which sends 1 to i_1, 2 to i_2 and so on, is $s = 0$ for the even permutations and $s = 1$ for the odd ones.

Example. An exterior 2-form ($k = 2$) is just an anti-symmetric bilinear form, $\omega(x, y) = -\omega(y, x)$, which can be defined by an antisymmetric matrix $\omega_{i,j} = -\omega_{j,i}$ (of order n for $V \approx \mathbb{R}^n$):

$$\omega(x_1e_1 + \cdots + x_ne_n, y_1e_1 + \cdots + y_ne_n) = \sum_{i,j} \omega_{i,j} \ x_iy_j \ .$$

---

*We recall the linearity definition

$$\omega(ax + by, z, \ldots, w) = a\omega(x, z, \ldots, w) + b\omega(y, z, \ldots, w)$$

for any $a \in \mathbb{R}$, $b \in \mathbb{R}$, for any $k + 1$ vectors $x, y, z, \ldots, w$ of V.

The addition of forms and the multiplication of forms by numbers are defined in the usual way, like for any function:

$$(\omega + \omega')(\xi_1, \ldots, \xi_k) = \omega(\xi_1, \ldots, \xi_k) + \omega'(\xi_1, \ldots, \xi_k) \ ,$$

for any k vectors $\xi_1, \ldots, \xi_k$ of the space V.

Theorem 3. *The set of the k-forms on a finite dimensional vector space V^n is a finite dimensional vector space.*

Its dimension and a basis of it are described below, in the proof.

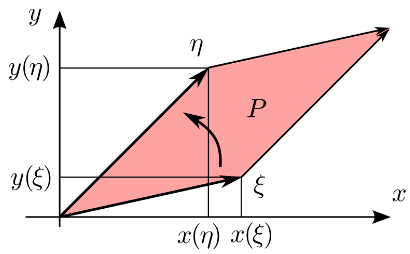
Example. For two vectors ξ, η in the oriented Euclidean plane $\mathbb{R}^2$ with Cartesian coordinates x and y, consider the (oriented) area ω of the parallelogram P generated by these vectors (Fig. 6.1),

$$P = \{a\xi + b\eta \in \mathbb{R}^2, 0 \leqslant a \leqslant 1, 0 \leqslant b \leqslant 1\} \ .$$

An elementary (and necessary) exercise shows, that

$$\omega(x, y) = \begin{vmatrix} x(\xi) & y(\xi) \\ x(\eta) & y(\eta) \end{vmatrix} = x(\xi)y(\eta) - y(\xi)x(\eta)$$

(which is the determinant, formed by the components of the vectors).



$$\mathrm{Area}(P) = \begin{vmatrix} x(\xi) & y(\xi) \\ x(\eta) & y(\eta) \end{vmatrix}$$

Figure 6.1: The oriented area of a parallelogram on the oriented Euclidean plane.

Remark. This calculation was the basic reason to introduce the determinants, which are the oriented volumes of the parallelepipeds generated by the lines (or by the columns) of the determinant in the oriented Euclidean space $\mathbb{R}^n$.

In the present day presentations of the algebraic theory of determinants, this basic geometric fact is always hidden from the students (to enhance the authority of the professors of algebra). All the theorems of the theory of determinants are in fact simple corollaries of this hidden geometric fact.

The preceding example provides an interesting volume n-form in $\mathbb{R}^n$. But it can also be used to construct interesting k-forms (for any $k < n$) in the n-dimensional vector space.

Example. Consider the three Cartesian coordinate functions

$$x : \mathbb{R}^3 \to \mathbb{R}\ , \quad y : \mathbb{R}^3 \to \mathbb{R}\ , \quad z : \mathbb{R}^3 \to \mathbb{R}\ .$$

The two functions x and y define a projection $p : \mathbb{R}^3 \to \mathbb{R}^2$ sending the initial 3-space onto a plane on which we will denote the coordinates by the same characters x and y, as for the coordinates in $\mathbb{R}^3$.

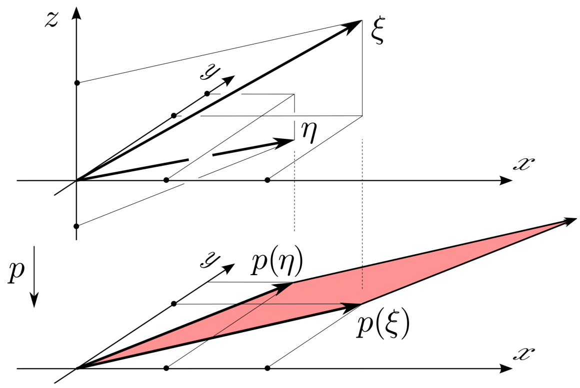


Figure 6.2: The 2-form $x \wedge y$ in 3-space, equal to the oriented area of the projected parallelogram.

A pair of vectors ξ, η of the space $\mathbb{R}^3$ is projected (Fig. 6.2) to the pair $p(\xi), p(\eta)$ of vectors of the plane $\mathbb{R}^2$. Associate to these two projected vectors the (oriented) area of the parallelogram generated by them in $\mathbb{R}^2$. This oriented area is an exterior 2-form of the original two vectors of the 3-space:

$$\omega(\xi, \eta) = \begin{vmatrix} x(\xi) & y(\xi) \\ x(\eta) & y(\eta) \end{vmatrix} = x(\xi)y(\eta) - y(\xi)x(\eta)\ .$$

Notation. This form is denoted by the symbol $x \wedge y$.

Similarly, the oriented k-volume of the k-parallelepiped, obtained in space $\mathbb{R}^k$ from k vectors $\xi_1, \ldots, \xi_k$ of the n-space $V \approx \mathbb{R}^n$ with coordinates $x_1, \ldots, x_n$ by the projection to the coordinate plane $(x_{i_1}, \ldots, x_{i_k})$ is an exterior k-form

$$\omega(\xi_1, \ldots, \xi_k) = \begin{vmatrix} x_{i_1}(\xi_1) & \ldots & x_{i_k}(\xi_1) \\ \vdots & & \vdots \\ x_{i_1}(\xi_k) & \ldots & x_{i_k}(\xi_k) \end{vmatrix}\ .$$

This k-form is denoted by the symbol

$$x_{i_1} \wedge x_{i_2} \wedge \cdots \wedge x_{i_k} .$$

Using the preceding examples, we shall prove the above theorem on the vector space structure of the set of k-forms on $V \approx \mathbb{R}^n$. The projections of the coordinate k-planes provide a lot of k-forms $x_{i_1} \wedge \cdots \wedge x_{i_k}$. These forms, however, are not linearly independent vectors of the space of the k-forms.

Indeed, $x \wedge x = 0$, and we should avoid any repetition in the sequence $(i_1, \ldots, i_k)$. We should also avoid the repetition of the same unordered set $\{i_1, \ldots, i_k\}$, since, say, the forms $x \wedge y$ and $y \wedge x$ are linearly dependent (because they have just opposite sign $x \wedge y = -y \wedge x$).

Lemma. *The k-forms $x_{i_1} \wedge x_{i_2} \wedge \cdots \wedge x_{i_k}$ in the vector space $V \approx \mathbb{R}^n$ corresponding to $i_1 < i_2 < \cdots < i_k$, are linearly independent and do form a basis of the vector space of the exterior k-forms on V.*

The dimension of the vector space of the exterior k-forms in a vector space of dimension n is equal to the number of combinations (*binomial coefficient*)

$$\binom{n}{k} = \frac{n(n-1)\ldots(n-k+1)}{1 \cdot 2 \cdot \cdots \cdot k} = \frac{n!}{k!(n-k)!} .$$

Proof. Let ω be an exterior k-form. Consider its values on the systems of k coordinate vectors,

$$\omega_{i_1,\ldots,i_k} = \omega(e_{i_1}, \ldots, e_{i_k}) \in \mathbb{R} .$$

Given k different vectors of V, $\xi_1, \ldots, \xi_k$, consider the k-form constructed as the linear combination

$$\alpha = \sum_{i_1 < i_2 < \cdots < i_k} \omega_{i_1,\ldots,i_k} \, x_{i_1} \wedge x_{i_2} \wedge \cdots \wedge x_{i_k}$$

of the volumes of all the C_n^k projections of the parallelepiped generated by the k vectors $\xi_1, \ldots, \xi_k$ (arguments of the resulting k-form α).

Then the k-forms α and ω have equal values on any system of k different vectors of the coordinate system. Thus by linearity,

$$\alpha(\xi_1, \ldots, \xi_k) = \omega(\xi_1, \ldots, \xi_k)$$

for any system of k vectors of V.

Therefore, the $\binom{n}{k}$ forms $x_{i_1} \wedge x_{i_2} \wedge \cdots \wedge x_{i_k}$ (where $i_1 < i_2 < \cdots < i_k$) generate the space of all k-forms in $V \approx \mathbb{R}^n$.

Their linear independence is also evident : if a linear combination of these $\binom{n}{k}$ forms,

$$\sum \alpha_{i_1,\dots,i_k} x_{i_1} \wedge \cdots \wedge x_{i_k} ,$$

vanishes identically, then its values on the sequences of coordinate vectors $(e_{i_1}, \dots, e_{i_k})$ vanish also. Hence the coefficients $\alpha_{i_1,\dots,i_k}$ (equal to these values) vanish too. This proves the linear independence of the $\binom{n}{k}$ forms. □

Notation. The vector space of exterior 2-forms on V is denoted by $V^* \wedge V^*$, of the 3-forms – by $V^* \wedge V^* \wedge V^*$, of the k-forms – by

$$\underbrace{V^* \wedge \cdots \wedge V^*}_{k \text{ times}} = \Lambda^k(V^*) .$$

We shall explain later the notation, explaining that the space of 2-forms is some kind of product of the space V^* of 1-forms on V with itself.

Thus, the preceding theorem provides the vector space isomorphism

$$\Lambda^k(\mathbb{R}^{n*}) \approx \mathbb{R}^N , \quad N = \binom{n}{k} .$$

Example. $\Lambda^0(V^*) \approx \Lambda^n(V^*) \approx \mathbb{R}$ for the n-dimensional vector space V.

Example. $\Lambda^k(V^*) \approx 0$ for $k > n$, since $C_n^k = 0$ for $k > n$: the only exterior 4-form in $\mathbb{R}^3$ takes the zero value for every 4 vectors, $\omega(x, y, z, w) \equiv 0$.

6.2 Algebra of Forms

The vector space of the forms is very important, but in order to satisfy the algebraists, one should define the *algebra of forms.* The main part of this construction is the operator of "exterior multiplication" of forms:

Exterior Product. The *exterior product* $\alpha \wedge \beta$ of an exterior k-form α on V with an exterior ℓ-form β on V is the exterior $(k+\ell)$-form $\alpha \wedge \beta$ on V, whose values are computed from the values of α and β in the following way:

$$(\alpha \wedge \beta)(\xi_1, \dots, \xi_{k+\ell}) = \sum (-1)^{s(i_1,\dots,i_k,j_1,\dots,j_\ell)} \alpha(\xi_{i_1}, \dots, \xi_{i_k}) \beta(\xi_{j_1}, \dots, \xi_{j_\ell}) .$$

The summation is taken along the set of all the decompositions of the set $\{1, \ldots, k+\ell\}$ into two disjoint subsets $\{i_1, \ldots, i_k\}$ and $\{j_1, \ldots, j_\ell\}$. The sign $(-1)^s$ equals $+1$, if the permutation of the set $\{1, \ldots, k+\ell\}$ in the order $\{i_1, \ldots, i_k, j_1, \ldots, j_\ell\}$ is even and to -1 if it is odd.

The resulting product is an exterior $(k+\ell)$-form: the linearity is evident, and the anti-symmetry follows from the fact that the permutations act on s in the right way (that is, the product of even permutations is even, of even and odd ones is odd, and so on).

Example. For $n = 2$, $k = \ell = 1$, the definition becomes

$$(\alpha \wedge \beta)(\xi, \eta) = \alpha(\xi)\beta(\eta) - \alpha(\eta)\beta(\xi) \ ,$$

according to the two decompositions of the set $\{1, 2\}$ into two parts: $\{i = 1\}$, $\{j = 2\}$ for the first decomposition, providing the plus sign, and $\{i = 2\}$, $\{j = 1\}$ for the second decomposition, providing the minus sign.

Now, we see that our notation $x \wedge y$ (on page 177) is explained by the fact that the oriented area of the projected parallelogram is the exterior product of the linear functions defining the projection.

Example. For $n = 3$, $k = 1$, $\ell = 2$, we have 3 decompositions of the set $\{1, 2, 3\}$:

$$\begin{array}{lll} \{i = 1\} \ , \ \ \{j_1 = 2, j_2 = 3\} \ ; & (1, 2, 3) \text{ even, sign “}+\text{” } ; \\ \{i = 2\} \ , \ \ \{j_1 = 1, j_2 = 3\} \ ; & (2, 1, 3) \text{ odd, sign “}-\text{” } ; \\ \{i = 3\} \ , \ \ \{j_1 = 1, j_2 = 2\} \ ; & (3, 1, 2) \text{ even, sign “}+\text{” } . \end{array}$$

In this case the general definition provides the formula of the exterior product of a 1-form α with a 2-form β,

$$(\alpha \wedge \beta)(\xi, \eta, \zeta) = \alpha(\xi)\beta(\eta, \zeta) - \alpha(\eta)\beta(\xi, \zeta) + \alpha(\zeta)\beta(\xi, \eta) \ .$$

For instance, this formula expresses the determinant (which is the oriented volume of the parallelepiped in oriented Euclidean 3-space V) as the exterior product of the first coordinate form $\alpha = x$ with the area of the projection on the complementary coordinate plane, $\beta = y \wedge z$. This is the usual “expansion” of the determinant along the first line.

This example explains our notation $x \wedge y \wedge z$ (on p. 179) for the oriented volume of the linear projection to the 3-space, defined by linear functions x, y and z on V.

Problem. Prove that the product

$$x \wedge (y \wedge z \wedge \cdots \wedge w)$$

is equal to the oriented volume of the parallelepiped

$$x \wedge y \wedge z \wedge \cdots \wedge w\,.$$

Theorem 4. *The exterior multiplication has the following properties*:

(i) distributivity: $\alpha \wedge (\beta + \gamma) = (\alpha \wedge \beta) + (\alpha \wedge \gamma)$;

(ii) graded-symmetry: $\alpha \wedge \beta = (-1)^{k\ell} \beta \wedge \alpha$ for any k-form α and ℓ-form β;

(iii) associativity: $(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma)$.

The distributivity follows from the distributivity of the multiplication of the ordinary numbers:

$$\alpha(\xi)\big((\beta + \gamma)(\eta)\big) = \alpha(\xi)\beta(\eta) + \alpha(\xi)\gamma(\eta)\ .$$

The graded-commutativity (or graded-symmetry) follows from the fact that, to permute a group of k elements with a directly subsequent group of ℓ elements, one should make $k\ell$ transpositions: the first element of the second group, j_1, should jump over the k elements of the first group preceding it, $(i_1, \ldots, i_k)$, then the second element of the second group, j_2, has to jump over $i_1, i_2, \ldots, i_k$ in this order, and similarly every element of the second group should jump over all the k elements of the first group.

The associativity of the exterior multiplication is the difficult part of the algebra of forms. However, for 1-forms it follows from the determinant expansion along the first line:

$$\alpha \wedge (\beta \wedge \cdots \wedge \gamma) = \alpha \wedge \beta \wedge \cdots \wedge \gamma\,.$$

Thus it suffices to prove the associativity for the products of monomials:

$$(\alpha \wedge \cdots \wedge \beta) \wedge (\gamma \wedge \cdots \wedge \delta) = \alpha \wedge \cdots \wedge \delta\,,$$

where $\alpha, \beta, \ldots, \gamma, \delta$ are 1-forms. In determinant theory this equality is called the "Laplace expansion of the determinant". To prove it, it suffices to compare the values of the product of k coordinate 1-forms with ℓ coordinate 1-forms and of $k + \ell$ coordinate 1-forms on the k, ℓ and $k + \ell$ basic vectors, and to check the signs. We leave this easy work to the reader.

Example. Consider the space $\mathbb{R}^{2n}$, equipped with the coordinates

$$(p_1, \ldots, p_n, q_1, \ldots, q_n) \ ,$$

and define there the exterior 2-form

$$\omega = (p_1 \wedge q_1) + (p_2 \wedge q_2) + \cdots + (p_n \wedge q_n) \ .$$

This form is called *the Darboux normal form of a symplectic structure*, and the coordinates are called the *Darboux coordinates*.

PROBLEM. Find the exterior product $\wedge^n \omega = \omega \wedge \cdots \wedge \omega$ (n times).

SOLUTION. For $n = 2$ we have:

$$\begin{aligned}
&[(p_1 \wedge q_1) + (p_2 \wedge q_2)] \wedge [(p_1 \wedge q_1) + (p_2 \wedge q_2)] = \\
&\quad = (p_1 \wedge q_1) \wedge (p_1 \wedge q_1) + (p_1 \wedge q_1) \wedge (p_2 \wedge q_2) + \\
&\quad + (p_2 \wedge q_2) \wedge (p_1 \wedge q_1) + (p_2 \wedge q_2) \wedge (p_2 \wedge q_2) \ .
\end{aligned}$$

Using the associativity and then the anti-symmetry, we get

$$(p_1 \wedge q_1) \wedge (p_1 \wedge q_1) = p_1 \wedge (q_1 \wedge p_1) \wedge q_1 = -(p_1 \wedge p_1) \wedge (q_1 \wedge q_1).$$

But the anti-symmetry implies the vanishing $q_1 \wedge q_1 = 0$, and hence

$$(p_1 \wedge q_1) \wedge (p_1 \wedge q_1) = 0 \ ,$$

as well as for the other pair: $(p_2 \wedge q_2) \wedge (p_2 \wedge q_2) = 0$.

Hence we get the formula

$$\omega \wedge \omega = 2 \ p_1 \wedge q_1 \wedge p_2 \wedge q_2 \ .$$

Similarly, for larger n one obtains the important formula

$$w^{\wedge n} := \wedge^n \omega = c \cdot p_1 \wedge q_1 \wedge p_2 \wedge q_2 \wedge \cdots \wedge p_n \wedge q_n$$

We leave to the reader the pleasure to compute the coefficient $c = c(n) \neq 0$.

A geometric fact resulting from this is that the volume element of the Darboux phase space is equal, up to a constant factor, to the n-th power of the symplectic structure ω. This equality is very important in symplectic geometry, mechanics and physics. It is called the Liouville theorem and it implies *the preservation of the volumes in phase space by the symplectomorphisms* (and hence by the phase flows of the Hamilton differential equations).

This preservation of the volumes, called also the *incompressibility* of the phase space, is the mathematical base of the statistical physics.

Example. The attractors are impossible for any Hamilton vector field, since near an attractor the flow transforms (in due time) its neighbourhood in the phase space into a smaller domain which cannot have the same phase volume $dp_1 \wedge \cdots \wedge dq_n$ as the initial neighbourhood – Fig. 6.3.

This fact makes extremely difficult all questions on the stability for the Hamiltonian systems (like for the 3-body problem or for the problem of the future of the planetary systems): Their stability cannot be stable, as it happens to the non-Hamiltonian systems with their attractors, due to the Liouville theorem on the preservation of volumes.

Figure 6.3: Equilibrium points.

The geometry of the Hamilton systems imposes also other restrictions on the phase flows. Say, an incompressible fluid can percolate from one side of a separating wall to the other side by any hole in the wall, whatever small it is.

A symplectic camel is unable to percolate from one halfspace (say, of $\mathbb{R}^4$) to the other, if the hole is sufficiently small. The camel is called "symplectic" in this theorem, since the "percolation" is supposed to preserve the symplectic form ω (as it happens when the phase flow is generated by a Hamiltonian vector field).

6.3 Integration of differential 1-forms

Returning to theory of integration on manifolds, we need to consider *differential forms*, that are *smooth fields of k-forms on the tangent spaces of the manifold at its different points.*

A differential k-form ω on a smooth manifold M^n is a family of algebraic exterior k-forms on the vector spaces T_xM^n: At each point $x \in M^n$ the corresponding algebraic k-form $\omega(x)$ belongs to the vector space

$$\Lambda^k(T_x^*M^n),$$

where $T_x^*M^n$ denotes the dual space V^* of the n-dimensional tangent space $V = T_xM^n$ of the manifold M^n at the point x.

Example. Let $M^n = \mathbb{R}^n$, with coordinates $(x_1, \ldots, x_n)$.

Definition. The 1-form dx_k is a linear function of the tangent vector v, associating to it its component v_k (in the fixed coordinate system).

Remark. The differential of the function x_k on M^n is not small at all: for a long vector v its component v_k is, in general, very large.

Every differential 1-form α on our manifold can be written at every point as the linear combination of the n special forms dx_k:

$$\alpha = a_1 dx_1 + \cdots + a_n dx_n \ ,$$

where the coefficients $a_1, \ldots, a_n$ may depend on the point x, being in general functions on $\mathbb{R}^n$.

In the general case of a curved manifold, one can write the same formulas locally, in a neighbourhood covered by the coordinate system. The coefficients a_k will be different in a different coordinate system, but the form itself is well defined independently of the choices of coordinates: it is a linear function of the tangent vector. The form is called smooth (analytic, continuous, ...) if the functions $a_k(x)$ which define the form, in the fixed coordinate system, are smooth (analytic, continuous, ...).

Now we are ready to start the integration theory.

Let γ be a smooth compact curve on the manifold V, and α a smooth differential 1-form (Fig. 6.4).

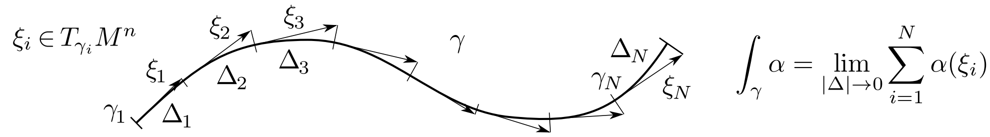


Figure 6.4: Definition of the integral of a form α along a curve γ.

The integral of the form α along the curve γ is defined by the following construction.

We subdivide the curve into small pieces Δ_i. We approximate each of them by a short tangent vector, ξ_i. We add the values of the form α at all these tangent vectors, and the integral is the limit of these sums for $|\Delta_j| \to 0$:

$$\int_\gamma \alpha = \lim_{\max |\Delta_j| \to 0} \left(\sum_{i=1}^{N} \alpha(\xi_i) \right) \ . \tag{1}$$

Example. For the curve γ defined by the parametrisation

$$x_k = \varphi_k(t) \ , \quad 1 \leqslant k \leqslant n \ , \quad 0 \leqslant t \leqslant 1 \ ,$$

we decompose the curve dividing it by the points $t_j = j/N$, and we approximate the parts Δ_j by the vectors $\xi_j = \varphi(t_{j+1}) - \varphi(t_j)$ with components

$$(dx_k)(\xi_j) = \left(\frac{d\varphi_k}{dt}\right)\Big|_{t_j} \left(dt = \frac{1}{N}\right) + o\left(\frac{1}{N}\right) \ .$$

The 1-form α has the expression $\alpha = a_1 dx_1 + \cdots + a_n dx_n$.

Therefore, the definition of its integral is

$$\int_\gamma \alpha = \lim_{N\to\infty} \sum_{j=0}^{N-1} \alpha\left(\frac{d\varphi}{dt}\Big|_{t_j} dt\right) = \sum_{k=1}^{n} \int_0^1 a_k\big(\varphi(t)\big)\frac{d\varphi_k}{dt}\ dt \ .$$

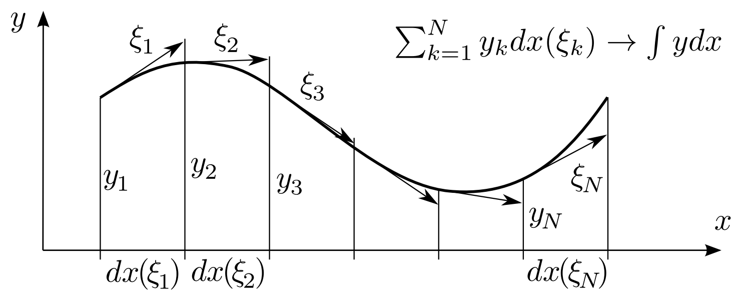


Figure 6.5: Integration of the differential form $\alpha = y\ dx$ along the graph of a function f.

For instance, the integral of the form $y\ dx$ along the graph $\{y = f(x)\}$ of the function f is the oriented area of the curvilinear trapezoidal domain of Fig. 6.5 (explaining the Leibniz notation $\int f(x)\ dx$ for this oriented area).

Remark. An extremely important property of the integration is the possibility to replace the summands in the defining formula (1) by other quantities for which the limit value does not change.

Namely, $\alpha(\xi_j)$ is a quantity of the first order of smallness, like Δ_j, when $|\Delta_j| \to 0$ in (1). Replacing these summands by different numbers, say, $\widetilde{\alpha}_j$, we shall not change the limit, if

$$|\widetilde{\alpha}_j - \alpha(\xi_j)| \leqslant C|\Delta_j|^2 \ , \tag{2}$$

since, being the sum $\sum |\Delta_j|$ bounded, the sum

$$\sum_j C|\Delta_j|^2 \leqslant C_1 \max |\Delta_j|$$

tends to 0 for $\max |\Delta_j| \to 0$ in (1).

Similarly, it suffices to have instead of (2) a weaker inequality

$$|\widetilde{\alpha}_j - \alpha(\xi_j)| \leqslant C|\Delta_j|\varepsilon \ ,$$

where $\varepsilon \to 0$ for $\max |\Delta_j| \to 0$.

6.3.1 Work against a Force Field

Consider a smooth curve γ, connecting two points γ_0 and γ_1 in Euclidean space $\mathbb{R}^n$ with Cartesian coordinates $x_1, \ldots, x_n$. Let F be a smooth differential 1-form on this space. It can be written as $F = F_1 dx_1 + \cdots + F_n dx_n$ for some functions F_k.

Work. The integral of the 1-form $-F$ along the curve γ is called the *work against the force field* F along the path γ (the functions $F_1(x), \ldots, F_n(x)$ are the components of the vector field F at the point x).

Problem. Find the work against the gravitational field with components $(0, 0, -g)$ in coordinates (x, y, z).

Answer. For $\gamma_0 = (x_0, y_0, z_0)$ and $\gamma_1 = (x_1, y_1, z_1)$ the work against the gravitational field along the path γ equals

$$\int_\gamma g \ dz = g(z_1 - z_0) \ .$$

Therefore, it does not depend on the form of the curve γ, depending just of its boundary points $\partial\gamma = \gamma_1 - \gamma_0$.

This independence of the integrating path is a particularity of our special field. Thus, for the form $y \ dx$ (and for the corresponding field with components $(y, 0)$ in Euclidean plane (x, y)) the work against the field is very different for different paths connecting the same endpoints (see Fig. 6.5).

Problem. Consider the vector field with components $(-y/(x^2+y^2), x/(x^2+y^2))$ on the Euclidean plane with deleted point 0 (x and y are Cartesian coordinates). Prove that the work against this field along a path lying in the complement to 0 in $\mathbb{R}^2$, and joining two fixed end-points, does not change under a homotopy of the path, being however different for non-homotopic paths.

Problem. Prove that the work against the gravitational field

$$F(X) = -\frac{X}{|X^3|}$$

in Euclidean space ($X \in \mathbb{R}^3 \smallsetminus 0$) along a path connecting two given points, is independent of the path connecting those points.

Answer. The function U in $\mathbb{R}^3 \smallsetminus 0$, defined by $U(X) = -1/|X|$, is the potential energy of the gravitational field $F(X) = -\text{grad } U = -X/|X|^3$. The work against the gravitational field along a path γ is equal to the increment of the potential energy U (and hence it does not depend on the path, depending only on the two boundary points, $\partial\gamma$, of the curve γ).

Problem. Find the average value of the potential energy U of the gravitational field along the surface of the sphere of radius R, centred at the point X.

Hint. The strange answer to this problem will be discussed on page 224. One interesting property of that answer is its insensitivity to the small deformations of the radius R, provided that $R \neq |X|$.

Remark (On the gradient). A function U on a Riemannian manifold* (for example, in Euclidean space) is called the *potential energy* for the vector field $F = -\text{grad } U$. The *gradient* of a function f on a Riemannian manifold is the vector field defined by the condition that, at every point, the differential of the function is equal to the scalar product with the vector of this field at that point:

$$df_x(\xi) := \langle \operatorname{grad} f \,,\, \xi \rangle_x$$

---

*A *Riemannian manifold* is a smooth manifold provided with a Riemannian metric on it (see the footnote of p.168). The geometry of 2-dimensional Riemannian manifolds is discussed in Ch. 8 (focusing on Lobachevsky plane) and for any dimension in Ch. 9.

for any vector ξ of the tangent space. In Euclidean Cartesian coordinates x_k the gradient field is defined by the components

$$\left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right) .$$

In bad calculus books this formula is the definition of the gradient vector field. It is, however, a wrong definition. The vector field provided by such components depends on the system of coordinates and has no geometrical meaning. The vector with components $(\partial f/\partial y_k)$, in the coordinate system $(y_1, \ldots, y_n)$, does *not* coincide with the vector with components $(\partial f/\partial x_k)$ in the initial coordinate system: these two tangent vectors are directed differently (for the same function f on the manifold).

The gradient vector field of a function depends on the Riemannian structure, and if the structure is fixed, it is a well-defined vector field which shows the direction of the fastest growth of the values of the function. However, the components of the vector grad f are expressed in terms of the partial derivatives of the function f by more complicated formulas, different from the expression $\partial f/\partial x_k$ of the wrong definition of the textbooks.

If a vector field of forces is minus the gradient of some potential energy U, that is, if $F = -\text{grad } U$, then the work against this field of forces along any path γ connecting an initial point γ_0 to a final point γ_1, is equal to the potential energy increment, $U(\gamma_1) - U(\gamma_0)$, and hence it is independent of the connecting path.

If the work against a vector field of forces is independent of the paths, then the field is the gradient of a well defined function whose value at γ_1 can be defined as the work along a path connecting γ_1 to a fixed point γ_0 (chosen arbitrarily).

Problem. Find the potential energy of the two-dimensional gravitational field,

$$F(X) = -\frac{X}{|X|^2}$$

(in Euclidean space $\mathbb{R}^n$ the standard gravitational field is $F(X) = -X/|X|^n$).

Calculate the mean value of the potential energy of the 2-dimensional gravitational field along a circle of radius R, centred at the point $X \in \mathbb{R}^2 \setminus 0$ (see p. 225 for the solution).

6.4 Integration of k-forms along k-submanifolds

Let us consider the multiple integrals. If we wish to integrate "something" along a (smooth) k-dimensional submanifold of a manifold M^n, the object that we shall integrate is, of course, a k-form ω defined along the tangent spaces of M^n: for a point $x \in M^n$ this form belongs to the space

$$\Lambda^k(T_x^* M^n)$$

and takes real values when it is evaluated on any k vectors of the real vector space $V = T_x M^n$ of dimension n.

Hence, if a local coordinate system is fixed, the form at x can be written as a linear combination of the basic forms,

$$\omega = \sum \omega_{i_1,\ldots,i_k}\, dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k}\,,$$

where $i_1 < i_2 < \cdots < i_k$ belong to $\{1, 2, \ldots, n\}$.

The only difference with the algebraic case is that this representation depends on the point x, that is, its coefficients $\omega_{i_1,\ldots,i_k}$ are functions of the point x (in the neighbourhood covered by the coordinate system).

We shall mostly consider *smooth exterior differential k-forms*, for which the above coefficients are smooth functions – say, of class C^r in some problems, C^∞ in most problems, but sometimes analytic or holomorphic.

The integral of a k-form is defined, as in the 1-dimensional case (Fig. 6.4, p. 184), as the limit of the sums of the values of the form on the small approximating parallelepipeds of the pieces $\Delta(j)$, into which we shall subdivide the k-dimensional manifold γ along which we integrate.

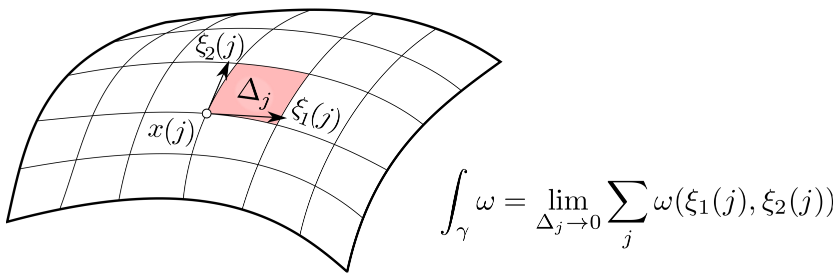


Figure 6.6: Definition of the integral of an exterior 2-form ω along a 2-dimensional oriented surface γ.

To construct a subdivision in a domain covered by a coordinate system $(t_1, \ldots, t_k)$ on γ^k, it suffices to take the coordinate hyperplane

$$\left\{ t_m = \frac{j_m}{N}\ ,\ j_m \in \{0, 1, \ldots, N-1\} \right\}$$

(see Fig. 6.6) and to take the limit at $N \to \infty$ of the sums of the values of the k-form ω,

$$\omega(\xi_1(j), \ldots, \xi_k(j))$$

on the tangent vectors $\xi_m(j) \in T_{x(j)}\gamma$, defined by the edges of the pieces $\Delta(j)$. Denote by $\gamma(t)$ the point of the manifold γ^k with coordinate values t. We may describe the point $x(j) = \gamma(t(j))$ by its coordinates $(t_1 = j_1/N, \ldots, t_k = j_k/N)$, where $j = (j_1, \ldots, j_k)$ is an integral vector belonging to the finite set $\{0, 1, \ldots, N-1\}^k$.

In these notations the tangent vector $\xi_m(j) \in T_{x(j)}\gamma$ is the vector

$$\left.\frac{\partial \gamma}{\partial t_m}\right|_{t(j)} (dt_m = 1/N) \ .$$

The existence of the limit for $N \to \infty$ is a standard exercise in calculus*, and we shall now use integrals of smooth differential forms ω on n-manifolds along k-dimensional oriented submanifolds γ,

$$\int_\gamma \omega \in \mathbb{R} \ .$$

We have used the coordinate system $(t_1, \ldots, t_k)$ to define the integral, but the resulting limiting value independence on the coordinate system is the well known theorem on the multiple integrals.

Example. Consider an oriented surface in the Euclidean 3-space and a vector field v with components (P, Q, R) in the Cartesian (orthonormal) coordinate system (x, y, z). The standard notation is

$$v = P\ \partial/\partial x + Q\ \partial/\partial y + R\ \partial/\partial z \ .$$

Here $\partial/\partial x$ denotes the standard coordinate vector field $e = \text{grad}\ x$ and so on, due to the fact that the derivative of any function u along this vector field e is its partial derivative $\partial u/\partial x$.

Now, we shall associate to each vector field v in our Euclidean space a differential 2-form ω_v, called "the flux of v" (Fig. 6.7).

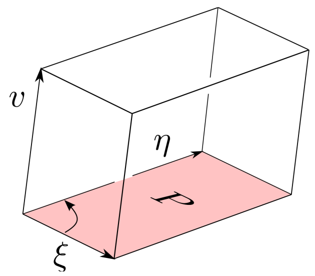
---

*As for integrals along curves, the summands $\omega(j) = \omega(\xi_1(j), \ldots, \xi_k(j))$ can be replaced by other quantities $\widetilde{\omega}(j)$, provided that the difference is of higher order of smallness than $\omega(j)$. Namely, in our notations here for the k-multiple integral of a differential k-form, the value $\omega(j)$ is a small quantity of order $1/N^k$ (since the number of the summands is of order N^k). Hence, it suffices that $\widetilde{\omega}$ verifies the inequality

$$|\widetilde{\omega}(j) - \omega(j)| \leqslant C(1/N)^{k+1} \quad \text{or even} \quad |\widetilde{\omega}(j) - \omega(j)| \leqslant C_1(1/N)^k \varepsilon \ ,$$

where $\varepsilon \to 0$ for $N \to \infty$.

This remark explains, among other things, the possibility of changes of variables in the multiple integrals: the subdivision into "parallelepipeds" depends on the coordinate system, but one evaluates the resulting correction (due to the difference of the small curvilinear parallelepipeds in both systems) to be a small quantity of higher order than the volumes of the parallelepipeds.



$$\omega_v(\xi, \eta) = (v, \xi, \eta)$$

Figure 6.7: The flux of a homogeneous liquid with velocity v, through the oriented parallelogram P in oriented Euclidean space $\mathbb{R}^3$.

The value of this form on two tangent vectors ξ, η of our space at some point a is equal to the triple product $[v(a), \xi, \eta]$, that is, to the oriented volume of the parallelepiped with edges $v(a)$, ξ and η. The physical meaning, explaining the word flux, is that the flow of a fluid with velocity v across the parallelogram generated by (ξ, η) fills with the new-coming fluid (in time unity) just this parallelepiped, provided that the velocity field is constant (that is, invariant by translations).

In the more realistic case in which the components of the field are variable, one should compute the local flux, replacing the vectors ξ and η by small vectors $\varepsilon\xi$ and $\varepsilon\eta$, and then dividing the resulting volume by ε^2 to get the local flux intensity $\omega_v(\xi, \eta)$ on finite vectors ξ, η, describing the fluxes across the infinitesimal parallelograms.

Problem. Find the flux of the gravitational field $v = -X/|X|^3$ in $\mathbb{R}^3$ through the sphere $|X| = R$ and write explicitly the 2-form ω_v.

Answer. The flux -4π is strangely independent of the radius R (the gravitational field in $\mathbb{R}^n$ has a similar property). The explicit formula for the form ω_v, where $v = P\ \partial/\partial x + Q\ \partial/\partial y + R\ \partial/\partial z$ is

$$P\ dy \wedge dz + Q\ dz \wedge dx + R\ dx \wedge dy$$

(memorise the nice cyclical order!).

Therefore, the flux of the gravitational field across a surface σ^2 is the double integral

$$I = \iint_{\sigma^2} -\frac{x\ dy \wedge dz + y\ dz \wedge dx + z\ dx \wedge dy}{(x^2+y^2+z^2)^{3/2}}\ , \quad \sigma^2 \subset \mathbb{R}^3 \setminus 0\ .$$

Problem. Prove that this integral does not change its value under a homotopy of the surface σ^2 inside the complement to the point 0 in $\mathbb{R}^3$.

Problem. Prove that this integral is zero, if the surface σ^2 is the boundary of a ball contained in the complement to the point 0 in $\mathbb{R}^3$.

Problem. Is it true that the integral I along a surface $\sigma^2 \in \mathbb{R}^3 \setminus 0$ whose boundary is a circle $\gamma = \partial\sigma^2$, is the same for all the surfaces whose boundary is the given curve γ?

Answer. No, it is in general false. However, it is true for those surfaces which are sufficiently close to σ^2.

The vector fields in a Euclidean space, whose flow along the surfaces bounding balls vanish, are called divergence-free fields. The gravitational field is divergence-free, the field $v(X) = X$ is not.

The divergence-free vector fields describe the fields of velocities of the incompressible fluids, for which the quantity of the fluid entering inside the boundary of the ball, in a given time interval, is equal to the quantity going out from it (through the other parts of the boundary).

As we shall see soon, a necessary and sufficient condition for this incompressibility property is

$$\operatorname{div} v = 0 \ ,$$

where the divergence of a vector field is the following function of the Cartesian orthonormal coordinates:

$$\operatorname{div} \left(P \,\frac{\partial}{\partial x} + Q \,\frac{\partial}{\partial y} + R \,\frac{\partial}{\partial z} \right) = \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z} \ .$$

The divergence is a natural generalisation of such notions as the speed of a motion, the density of the matter distribution, the Gaussian curvature of a surface in Euclidean space and so on.

To explain all these physical and geometrical notions (and many others), we shall introduce the general notion of the *exterior derivatives of differential forms* which provides a powerful understanding of the integration of the differential forms on the smooth manifolds. But before that, we will introduce the "chains": the domains along which one integrates the differential forms.

6.5 Chains and boundaries

Let σ^m be a small piece of an m-dimensional submanifold of manifold M^n, where a smooth differential m-form is defined.

Example. For $m = 1$, σ^1 is just a small segment of curve, bounded by two end-points, γ_0 and γ_1.

We shall denote the boundary hypersurface $\partial\sigma^m$ by γ^{m-1}. We suppose that the domain σ^m is oriented, and we orient γ^{m-1} as the boundary of σ^m.

It means that the orienting tangent frame of the boundary manifold, preceded by the exterior normal vector, orients the manifold σ^m positively (Fig. 6.8 a).

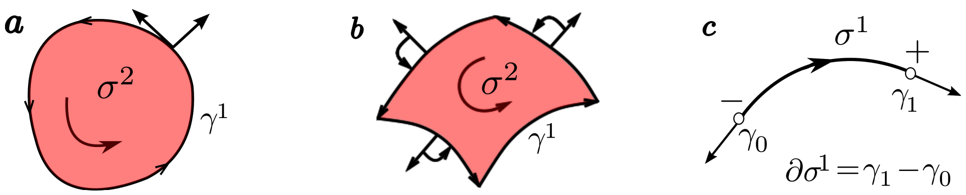


Figure 6.8: The orientation of the boundary of an oriented "manifold".

No orientation of the ambient manifold M^n is used here, and M may be non oriented and even non-orientable.

The condition of the smoothness of the boundary may be relaxed. For instance, the piecewise smoothness, as in Fig. 6.8 b, would suffice. In such cases one calls the resulting generalised manifold a "chain".

More formally, to define chains we have first to consider oriented m-dimensional "curvilinear polyhedra" in M^n (called singular polyhedra –Fig. 6.9), which are the "bricks" from which the chains are constructed.

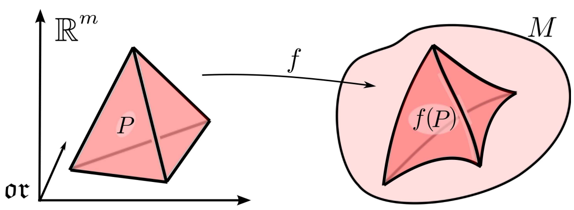


Figure 6.9: A singular m-dimensional polyhedron.

Definition. A *singular m-dimensional polyhedron* σ of M^n is a smooth map of an oriented convex polyhedron P in Euclidean space $\mathbb{R}^m$ to M^n.

In other words, it is a triple $\sigma = (P, f, \mathfrak{or})$ in which $P \subset \mathbb{R}^m$ is a convex polyhedron, $f : P \to M^n$ is a differentiable map and $\mathfrak{or}(P)$ is the choice of an orientation of the polyhedron P, that is, of the space $\mathbb{R}^m$. The singular polyhedron obtained from σ by taking the opposite orientation is denoted by $-1 \cdot \sigma$ (or $-\sigma$) and called the *negative of* σ.

Integrating along curves it is often useful to consider contours consisting of several pieces, having the possibility to traverse them any number of times and in either direction. The higher dimensional contours along which we shall integrate the differential forms are the *chains*.

Definition. A *chain* c^m of dimension m in M^n is a formal sum of a finite collection $\sigma_1^m, \dots, \sigma_r^m$ of m-dimensional singular polyhedra, taken with integral multiplicities $n_1, \dots, n_r$. That is, $c^m = n_1\sigma_1^m + \cdots + n_r\sigma_r^m$.

For $m = 0$ we have to precise that a *zero-dimensional chain* is a collection of points with multiplicities, and the "orientation" of a point is given by a sign $+$ or $-$ (this is the zero-dimensional version of an oriented frame).

We can think a chain as formed by "bricks": the singular polyhedra entering in the definition are the bricks of the chain.

Example. The boundary of σ^2 in Fig. 6.8 *b* is an example of chain. It is the sum of the four oriented smooth parts of γ^1 taken with multiplicity 1.

Example. Any sum of the oriented m-dimensional faces of a polyhedron with integral multiplicities is a chain (for example, the edges of a tetrahedron).

Remark. In the above definition, it is possible to restrict the convex polyhedrons P to be m-dimensional cubes or m-dimensional simplices (leading to the "cubic chains" or "simplicial chains"), the same geometric "generalised manifold" may be represented as a chain in any of these special senses.
The coefficients n_j in the definition of the chains are usually integers, but sometimes one uses real coefficients, or complex coefficients, or even mod p residues ($\mathbb{Z}_p$-coefficient chains), the case $p = 2$ being especially important for the "non-oriented chains".

The addition of chains is defined in the direct way

$$(n_1\sigma_1+\cdots+n_r\sigma_r)+(m_1\sigma_1'+\cdots+m_s\sigma_s') = n_1\sigma_1+\cdots+n_r\sigma_r+m_1\sigma_1'+\cdots+m_s\sigma_s' .$$

Group structure. *Using the natural identifications*

$$n_1\sigma_1 + n_2\sigma_2 = n_2\sigma_2 + n_1\sigma_1, \quad n\gamma + m\gamma = (n+m)\gamma,$$

$$0 \cdot c^m = 0 \quad and \quad c^m + 0 = c^m \,,$$

the set $\{c^m\}$ *of* m*-chains on* M *acquires a structure of commutative group.*

This group has infinitely many generators (as many as m-dimensional singular polyhedrons $\sigma : P \to M$ exist). But for polyhedra we have:

Example. The *group of* m*-faces* of a polyhedron P is the set of all possible sums of the m-faces of P with integral multiplicities (linear combinations). It is isomorphic to the free abelian group $\mathbb{Z}^{|f_m|}$, where $|f_m|$ is the number of m-dimensional faces of P; each m-face F_j, taken with coefficient 1, being a generator. For instance, the groups of faces, of edges and of vertices of a tetrahedron are isomorphic to the respective groups $\mathbb{Z}^4$, $\mathbb{Z}^6$ and $\mathbb{Z}^4$; while those of the dodecahedron are isomorphic to the groups $\mathbb{Z}^{12}$, $\mathbb{Z}^{30}$ and $\mathbb{Z}^{20}$.

Oriented boundary of a chain. To define the boundary of a chain and formalise its induced orientation, we have first to do that for a polyhedron.

Given a convex oriented m-dimensional polyhedron P in Euclidean space $\mathbb{R}^m$, the *boundary* of P is the $(m-1)$-chain ∂P on $\mathbb{R}^m$ whose "bricks" γ_j (singular $(m-1)$-polyhedra) are the $(m-1)$-dimensional faces F_j of the polyhedron P together with maps $\varphi_j : F_j \to \mathbb{R}^m$ that embed the faces in $\mathbb{R}^m$ and with the orientations $\mathfrak{or}_j$; all multiplicities are equal to 1 :

$$\partial P = \sum \gamma_j \qquad \gamma_j = (F_j, \varphi_j, \mathfrak{or}_j)\,.$$

Consider an oriented frame $\mathbf{e}_1, \ldots, \mathbf{e}_m$ in $\mathbb{R}^m$. We choose an interior point of one of the faces F_j of the polyhedron P and there take a vector $\mathbf{n}$ normal to P in the outward direction. We say that a frame $\mathbf{f}_1, \ldots, \mathbf{f}_{m-1}$ of the face F_j *orients correctly* this face if the frame $\mathbf{n}, \mathbf{f}_1, \ldots, \mathbf{f}_{m-1}$ orients $\mathbb{R}^m$ in the same way as the frame $\mathbf{e}_1, \ldots, \mathbf{e}_m$ (Fig. 6.10).

For an oriented segment in $\mathbb{R}$ $(m = 1)$ each boundary point is *oriented* by giving it the sign $+$ or $-$ according as, at that point, the vector pointing out from the segment orients positively the segment or not. So the boundary of the oriented segment $\overrightarrow{F_1F_2}$ is $\partial(\overrightarrow{F_1F_2}) = 1 \cdot F_2 - 1 \cdot F_1$ (Fig. 6.10).

For $m = 0$, we assume the boundary of a point is empty.

The boundary of the bricks that form a chain (that is, of its singular polyhedra) is defined in the same way :

The *boundary* $\gamma = \partial\sigma$ *of an* m*-dimensional singular polyhedron* on M, $\sigma = (P, f, \mathfrak{or})$, is the $(m-1)$-dimensional chain $\partial\sigma = \sum \gamma_j$ whose brick

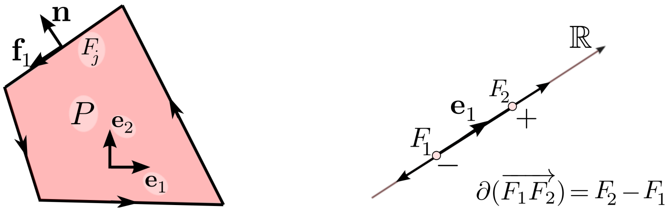


Figure 6.10: Orientation of the boundary of a polyhedron.

$\gamma_j = (F_j, f_j, \mathfrak{or}_j)$ is given by the $(m-1)$-dimensional face F_j of P, the restriction f_j of the map $f : P \to M$ to F_j and the above orientation $\mathfrak{or}_j$.

The *boundary of an m-dimensional chain* c^m on M is the $(m-1)$-dimensional chain on M defined as the sum of the boundaries of the bricks of c^m taken with their multiplicities:

$$\partial c^m = \partial(n_1\sigma_1 + \cdots + n_r\sigma_r) = n_1\partial\sigma_1 + \cdots + n_r\partial\sigma_r.$$

For example, in Fig. 6.11 we show the boundaries of the chains $A+B$ and $A-B$. Note that to follow a closed path along $\partial(A-B)$ one has to pass twice along the edge marked with double arrow (it has multiplicity 2 in the chain $\partial(A-B)$).

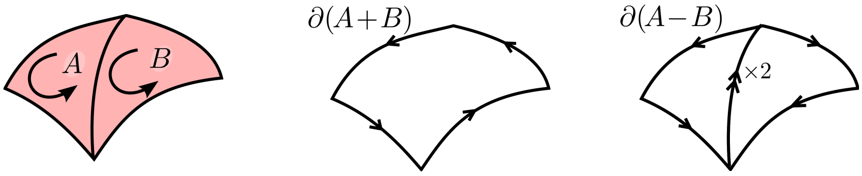


Figure 6.11: The boundaries of two chains.

Example. The boundary of a zero-dimensional chain is always zero.

The boundary homomorphism. Observe that the operation of taking the boundary of the m-dimensional chains of a manifold M, defines a map $\partial_m : \{c^m\} \to \{c^{m-1}\}$, $c^m \mapsto \partial c^m$. By the above definition of the boundary of a chain, this map is a homomorphism of abelian groups. It is called *boundary operator* or *boundary homomorphism.* For the polyhedra we also have:

Example. For a polyhedron P, the boundary homomorphism ∂_m associates to each linear combination of the oriented m-faces of P a linear combination of its oriented $(m-1)$-faces. For instance, taking the oriented boundaries in the groups of faces, of edges and of vertices of a tetrahedron we get three boundary homomorphisms: $\partial_f : \mathbb{Z}^4 \to \mathbb{Z}^6$, $\partial_e : \mathbb{Z}^6 \to \mathbb{Z}^4$ and $\partial_v : \mathbb{Z}^4 \to 0$. And, similarly, for a dodecahedron the three boundary homomorphisms are: $\partial_f : \mathbb{Z}^{12} \to \mathbb{Z}^{30}$, $\partial_e : \mathbb{Z}^{30} \to \mathbb{Z}^{20}$ and $\partial_v : \mathbb{Z}^{20} \to 0$.

The oriented boundary of the oriented boundary of a chain vanishes. That is, the composition of the boundary homomorphisms has the semi-exactness property $\partial_{m-1} \circ \partial_m = 0$ $(\operatorname{Im} \partial_m \subset \operatorname{Ker} \partial_{m-1})$. The geometric proof of this fact for convex polyhedra is given below (p. 211).

6.6 Exterior derivatives and Stokes Theorem

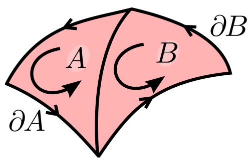
Observe that a function is a 0-form, and its differential is a 1-form. In order to define the exterior derivative of a form, we shall imitate the definition of the differential of a smooth function.

For any given smooth differential m-form ω we shall define a remarkable differential $(m+1)$-form Ω.

We start from any $(m+1)$-dimensional oriented piece σ^{m+1} with its oriented boundary $\partial\sigma^{m+1} = \gamma^m$ (Fig. 6.8) and we integrate the m-form ω along the boundary γ^m, defining in this way a function on the $(m+1)$-dimensional pieces,

$$F(\sigma^{m+1}) = \int_{\gamma^m = \partial\sigma^{m+1}} \omega \ .$$

This function is an additive function of the chain σ^{m+1} (Fig. 6.12), and contracting this chain to a very small $(m+1)$-dimensional "parallelepiped" in the tangent space to M^n, we naturally conclude that, up to a higher order correction, it is a multilinear and anti-symmetric function of $m+1$ tangent vectors of M at the same point.



$$\sigma = A + B$$

$$\int_{\partial\sigma} \omega = \int_{\partial A} \omega + \int_{\partial B} \omega$$

Figure 6.12: The additivity of the integrals of forms.

Denote by $\sigma^{m+1}(s_1, \dots, s_{m+1})$ the $(m+1)$-dimensional domain defined in M^n as the image of the parallelepiped $(0 \leqslant t_1 \leqslant s_1, \dots, 0 \leqslant t_{m+1} \leqslant s_{m+1})$ under the map $\sigma^{m+1} : \mathbb{R}^{m+1} \to M^n$ which defines the chain σ^{m+1} (Fig. 6.13).

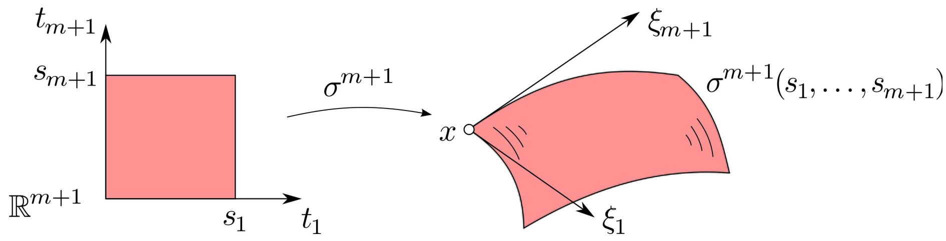


Figure 6.13: The elementary parallelepiped $\sigma^{m+1}(s)$ and the tangent vectors ξ_j.

Principal Part Theorem. *The Taylor series of the function* $F(s_1, \dots, s_{m+1})$ *has the form*

$$F(s_1, \dots, s_{m+1}) = \Omega \; s_1 s_2 \dots s_{m+1} + O(|s|^{m+2}) \; .$$

Generically, the coefficient Ω is different from 0, and in this case the first nonzero term of the Taylor series of the function F is multilinear of degree $m+1$.

Denote the point $\sigma^{m+1}(0)$ by x and denote by ξ_j the partial derivatives of the map $\sigma^{m+1} : \mathbb{R}^{m+1} \to M^n$ at this point:

$$\xi_j = \frac{\partial \sigma^{m+1}}{\partial t_j}\bigg|_{t=0} \in T_x M^n \quad (j = 1, \dots, m+1) \; .$$

Addition to Principal Part Theorem. *The coefficient* Ω *of the Principal Part Theorem depends only on the vectors* ξ_j, *and depends on these vectors as a multilinear anti-symmetric form:*

$$\Omega = \Omega(\xi_1, \dots, \xi_{m+1}) \; , \quad \Omega \in \underbrace{V^* \wedge \dots \wedge V^*}_{m+1} \; , \quad V = T_x M^n \; .$$

Geometrically this additional statement is an easy corollary of the additivity of F with respect to the piece σ^{m+1}, but we shall rather provide a less geometrical explicit coordinates-dependent calculation of the $(m+1)$-form Ω. The formula for the value of Ω, provided by these calculations, implies also the above more geometrical Addition.

These formulas are also practically applicable: Maxwell (who gave them the name "Stokes Lemma") based on these formulas his electromagnetic field theory, and later Bourbaki, trying to provide a nice proof of these formulas, created all their "Elements of Mathematics" books to prepare the Stokes Lemma proof (which they never arrived to write, as far as we know, and which had been the base of the Poincaré theory of multiple integrals, explained later to Bourbaki by E. Cartan and generalised by de Rham and by Whitney).

Calculation of the Principal Part of the integral along the boundary of a small piece. The following calculation generalises, for any m, the well known basic formula for $m = 0$:

$$f(s) - f(0) = \left(\frac{df}{ds}\right)_0 s + o(s) ,$$

where the left hand side "increment" is the "integral" along the 0-dimensional boundary ("s"–"0") of the $(0+1=1)$-dimensional chain $\{t : 0 \leqslant t \leqslant s\}$.

To explain the higher dimensional version, we consider the case $m = 1$.

Proposition. *The integral of a differential* 1*-form*

$$\omega = P(x,y)\ dx + Q(x,y)\ dy$$

along the boundary of a rectangle $\sigma^2 = \{0 \leqslant x \leqslant a\ ,\ 0 \leqslant y \leqslant b\}$, *of the oriented plane* $\mathbb{R}^2$ (Fig. 6.14) *has the following expression*

$$\int_{\partial\sigma^2} P dx + Q dy = \Omega\ ab + o(|a|^2 + |b|^2)\ . \tag{3}$$

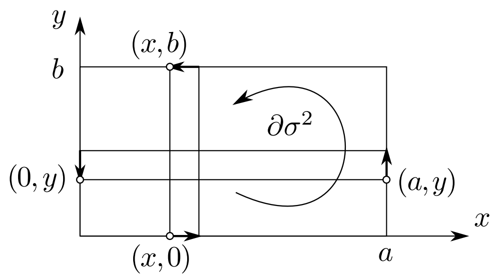


Figure 6.14: Calculation of the principal part of the integral of a 1-form along the boundary of a small domain.

Proof. Comparing the contributions of the boundary points $(x, 0)$ and (x, b), we get their contributions to the integral (3):

$$\int_0^a \big(P(x,0) - P(x,b)\big)\ dx = \int_0^a \left(\frac{\partial P}{\partial y}\ (-b) + O(b^2)\right)\ dx\ .$$

Similarly, comparing the contributions of the two boundary points $(0, y)$ and (a, y) we get the contribution

$$\int_0^b \big(Q(a,y) - Q(0,y)\big)\ dy = \int_0^b \left(\frac{\partial Q}{\partial x}\ a + O(a^2)\right)\ dy\ .$$

Replacing $\partial P/\partial y$ and $\partial Q/\partial x$ at the preceding integrals by their values at $x = y = 0$, we add an error of order $O(a^3 + b^3)$.

Adding the contributions of the horizontal and of the vertical parts of the boundary, we obtain the final expression for the integral (3),

$$\int_{\partial\sigma^2} Pdx + Qdy = \int_0^a \int_0^b \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\ dx\ dy + O(a^3 + b^3)\ ,$$

where $\partial Q/\partial x - \partial P/\partial y$ means the value of Ω on this combination of the partial derivatives at the origin.

Thus, integrating the constant Ω along the parallelogram, we obtain the result (3), together with the explicit formula

$$\Omega = \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\ ,$$

which also proves (in the case $m = 1$) that the number Ω depends on the tangent vectors ξ_1 and ξ_2 as a differential form of degree 2.

Namely, for $\omega = \sum_{i=1}^n f_i\ dx_i$ the same calculation provides the value

$$\Omega = \sum_{i=1}^n df_i \wedge dx_i = \sum_{j<i} \left(\frac{\partial f_i}{\partial x_j} - \frac{\partial f_j}{\partial x_i}\right) (dx_j \wedge dx_i)$$

for the coefficient of the principal part. In the preceding formulas, this 2-form was evaluated on the two tangent vectors $\xi_1, \xi_2 \in T_x M^n$ that generate the "small parallelogram" σ^2. □

The differential 2-form Ω, called *exterior derivative* of the 1-form ω (it is noted $\Omega = d\omega$), is the principal part of the integral of the original 1-form along the boundary of a small piece of a surface. As a corollary we get

Stokes Theorem (on the derivative of a 1-form). *For a smooth differential* 1*-form* ω *on* M^n *its integral along the oriented boundary of a* 2*-chain* σ^2 *is equal to the integral of the derivative of the* 1*-form* ω *along the chain* σ^2:

$$\int_{\partial\sigma^2} \omega = \int_{\sigma^2} d\omega \ .$$

Proof. Due to the linearity of the integrals with respect to the chains, it suffices to prove it for an elementary chain, say, for a small square map σ^2.

Subdividing the square into N^2 smaller squares, we represent the integral along the boundary of our square as the sum of the N^2 integrals of ω along the boundaries of the small squares.

Each of these integrals of ω is equal to the integral of the derivative form Ω along the interior small square, up to a correction whose order is at least cubic with respect to $1/N$.

Such corrections are neglectable for the calculation of the integral (see pages 185, 190). So, from the approximate Principal Part Theorem, we get

$$\int_{\partial\sigma^2} \omega = \int_{\sigma^2} d\omega \ . \qquad \square$$

Remark. The particular case of Stokes formula in which ω is a 1-form in three-dimensional oriented Euclidean space is of special interest, due to many applications in Physics. Namely, using the Cartesian orthonormal orienting coordinates x, y, z of Euclidean 3-space, the integral of the 1-form

$$\omega = P\ dx + Q\ dy + R\ dz$$

is the work of the corresponding vector field

$$v = P\ \partial/\partial x + Q\ \partial/\partial y + R\ \partial/\partial z$$

(page 186), along the 1-chain γ.

The derivative 2-form

$$\begin{aligned} d\omega &= dP \wedge dx + dQ \wedge dy + dR \wedge dz = \\ &= \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)(dx \wedge dy) + \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right)(dy \wedge dz) + \\ &+ \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right)(dz \wedge dx) \end{aligned}$$

can be represented as the flux of the vector field

$$w = \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right)\frac{\partial}{\partial x} + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right)\frac{\partial}{\partial y} + \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\frac{\partial}{\partial z}\ , \qquad (4)$$

(see page 190) which is called the *curl* of v or the *rotor* of v:

$$w = \mathrm{rot}\ v = \mathrm{curl}\ v\ .$$

A very useful "mnemonic" rule to memorise the long formula (4) is to write it in the form of a determinant

$$w = \begin{vmatrix} \mathrm{i} & \mathrm{j} & \mathrm{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ P & Q & R \end{vmatrix}\ ,$$

where i, j, k are the three basic vectors we have denoted above by $\partial/\partial x$, $\partial/\partial y$ and $\partial/\partial z$.

In these terms the preceding theorem takes the form:

Stokes Theorem. *The circulation*[*] *of a smooth vector field along the boundary of an oriented surface is equal to the flux of its curl across the bounded surface.*

Problem. Consider a rigid body, rotating in Euclidean oriented 3-space with angular velocity ω. The vector field of the velocities of the points of the body is provided by the vector-product,

$$v(x) = [\,\overrightarrow{\omega}, \overrightarrow{x}\,]\,.$$

Calculate the rotor of this vector field.

Answer. $\mathrm{rot}[\,\overrightarrow{\omega}, \cdot\,] = 2\omega$.

This example explains the origin of the word "rot" (from the word "rotation"): one should interpret the given vector field v, as providing to the medium some diminishing or growing (represented by the symmetric part of the derivative matrix) and also some rigid rotation (represented by the antisymmetric part, which can be written as $[\,\overrightarrow{\omega}, \cdot\,]$).

In this sense the rotor of a vector field provides the best rotational approximation of the motion of the particles, produced by that field.

[*] We recall that, given a vector field v on a Riemannian manifold, the circulation of v along a curve is defined as the integral along that curve of the 1-form ω defined by the condition that its value $\omega(\xi)$ on a tangent vector ξ at a point x is the scalar product of the vectors ξ and $v(x)$: $\omega(\xi) := \langle\, \xi\,,\, v(x)\,\rangle_x$.

Remark. The factor 2 in the above answer is due to the fact that the length of the circle of radius 1 is 2π, and the area of the disc that it bounds is π.

Remark. The "Stokes Theorem" (or "Stokes Lemma") was neither invented nor proved by Stokes. Its author was sir W. Thomson, lord Kelvin, who wrote a letter on it to Stokes.

Preparing the Cambridge written examination "Tripos", the Trinity College asked Stokes to provide a problem for it. Having nothing ready, he proposed the Kelvin statement as an examination problem.

Maxwell, going through this examination as a student, solved that problem. He was so impressed by it that he asked the administration who had proposed it. They gave him the Stokes name, and Maxwell called this important fact by the name of Stokes (discovering a lot of very useful applications of this theorem in physics). In mathematics, the basis of Cohomology theory was developed from the "Stokes Theorem" by Poincaré and later by Kolmogorov.

Explaining his cohomology theory, published in the Comptes Rendus de l'Académie des Sciences, Paris (1935), Kolmogorov mentioned that the main base and inspiration for his combinatorial topological construction came rather from physics than from algebra, his cohomologies were a generalisation of such notions of hydrodynamics of the incompressible fluids and of the Maxwell magneto-hydrodynamics, as the flux of fluid across a surface and the magnetic flux. Kolmogorov, as his mathematical predecessors, mentioned especially the Günter theory of "functions of domains" (called today "generalised functions" or "distributions" – see below pp. 206-208).

The exterior derivative of an m-form ω is an $(m+1)$-form Ω. It is the principal part of the integral of ω along the boundary $\partial\sigma^{m+1}$ of an oriented "small parallelepiped" σ^{m+1} generated by $m+1$ vectors of the tangent space to the ambient manifold M^n at the same point $x \in M^n$:

$$\int_{\partial\sigma^{m+1}} \omega = \Omega(\xi_1, \dots, \xi_{m+1}) + o(|\xi|^{m+1}) \ .$$

The proof (providing also the explicit formula for Ω) follows the same lines as that given for $m = 1$ (on page 199). For

$$\omega = \sum f_{i_1,\dots,i_m}(dx_{i_1} \wedge \cdots \wedge dx_{i_m})$$

we get

$$\Omega = \sum df_{i_1,\dots,i_m} \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_m} \ .$$

Example. Consider the flux 2-form ω of a vector field v in oriented Euclidean 3-dimensional space, which (in orthonormal oriented Cartesian coordinates x, y, z) are given by

$$\begin{aligned} v &= P\ \partial/\partial x + Q\ \partial/\partial y + R\ \partial/\partial z\ , \\ \omega &= P\ dy \wedge dz + Q\ dz \wedge dx + R\ dx \wedge dy\ . \end{aligned}$$

In this case

$$\Omega = (\partial P/\partial x + \partial Q/\partial y + \partial R/\partial z) dx \wedge dy \wedge dz \ ,$$

that is, in this case the derivative is the product of the (oriented) volume element $dx \wedge dy \wedge dz$ by a function. This function is called *the divergence* of the given vector field v:

$$\operatorname{div}(P\ \partial/\partial x + Q\ \partial/\partial y + R\ \partial/\partial z) = (\partial P/\partial x + \partial Q/\partial y + \partial R/\partial z) \ .$$

Similarly, the exterior derivative of the $(n-1)$-form ω_v which represents the flux of a smooth vector field on oriented Euclidean space $\mathbb{R}^n$, given in Cartesian orthonormal coordinates by

$$v = \sum v_i\ \partial/\partial x_i$$

is the n-form

$$d\omega_v = (\operatorname{div}\ v)\ \tau\ , \quad \tau = dx_1 \wedge \cdots \wedge dx_n\ ,$$

where div $v = \sum_{i=1}^n (\partial v_i/\partial x_i)$.

Remark. The physical meaning of the divergence of a vector field is the velocity of growing of the volumes, provided by its flow.

Problem. Suppose the basic unit vectors $(e_1, \ldots, e_n)$ of Euclidean n-space are slightly deformed to become $(e_k + tv_k)$ $k = 1, \ldots, n$, $|t| \ll 1$. Calculate the volume enlargement of the unit cube, in first order with respect to t, in terms of the components of the deforming vectors

$$v_k = \sum_{\ell=1}^{n} v_{k,\ell}\ e_\ell \qquad (k = 1, \ldots, n)\ .$$

Answer. $V = 1 + at + O(t^2)$, where $a = \sum_{k=1}^n v_{k,k}$ is the trace of the deformation matrix:

$$\frac{dV}{dt} = \operatorname{tr}\ (v_{k,\ell}) = \operatorname{div} \left(\sum_{\ell=1}^{n} w_\ell\ \partial/\partial x_\ell \right)\ ,$$

where $w_\ell = \sum v_{k,\ell}\ x_k$ is the linear vector field, deforming the vertices e_k of the cube with velocities v_k.

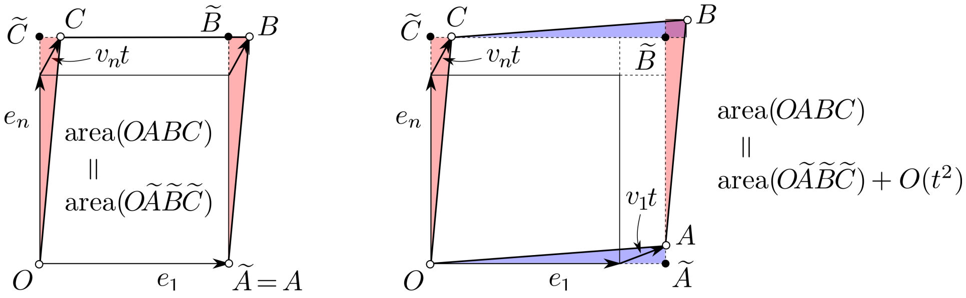


Figure 6.15: Geometry of traces and of divergences: the increment of the volume of the cube depends (on first approximation) only on the deformation of each edge in its own direction.

This result, showing also the importance of the traces, has an elementary geometric explanation: for a small deformation of the edges of a parallelepiped only the deformation along the edge itself is relevant (in first approximation), the deformations, directed along all the other vertices, contributing only higher order changes of the volume (Fig. 6.15).

The Stokes theorem follows from the expression of the principal part by the same reasoning that in the case $m = 1$ (page 202):

Stokes Theorem. *The integral of any smooth differential m-form ω along the boundary of an oriented $(m+1)$-chain σ^{m+1} is equal to the integral of the $(m+1)$-form $d\omega$ along the bounded chain σ^{m+1}:*

$$\int_{\partial\sigma^{m+1}} \omega = \int_{\sigma^{m+1}} d\omega \ . \tag{5}$$

Taking this theorem into account, we can interpret the integral of an $(n-1)$-form ω_v along the boundary of an n-domain σ^n in the oriented Euclidean n-space as the quantity of fluid (whose field of velocities is v) produced in the unity of time inside the domain σ^n. In this sense the divergence of a vector field is the density of the production of fluid by the distributed "sources": the flux of fluid through the boundary of a domain is equal to the integral of the sources density inside that domain.

6.7 “Functions of domains” and δ-function

Problem. Calculate the divergence of the gravitational field $F(X) = -\dfrac{X}{|X|^n}$ in Euclidean n-dimensional space.

Answer. Calculating the partial derivatives, one could get that div $F = 0$, but it is wrong.

Solution. Indeed, to calculate the flux across the hypersphere of radius R centred at the origin, we have to integrate the function $-1/R^{n-1}$ (which is the normal component of the field) along the sphere of radius R in $\mathbb{R}^n$. Denote the $(n-1)$-volume of the hypersphere of radius 1 by $c(n)$. We get the flux

$$\int_{|X|=R} -R^{1-n}\ d\sigma = -c\ .$$

Hence the integral of the divergence of the field along the ball $|X| \leqslant R$ is equal to $c \neq 0$. For $n = 3$ we know the value $c(3) = 4\pi$.

The rotational symmetry of the field implies that its divergence is also symmetric by rotations: div $F = f(R)$.

The independence of the flux from the radius of the sphere shows that $f = 0$ for any $R > 0$.

Thus, the mysterious function f should be zero everywhere (outside the point 0 at which the field is not smooth and the divergence value is not defined), while its integral along the ball $|X| \leqslant R$ is equal to $-c \neq 0$.

Such functions do not exist in mathematics, and hence physicists invented new names for them.

In the time of the 1st world war, N. Günter, a mathematician of Saint Petersburg, created the “theory of functions of domains”, whose integrals along the domains are well defined, but their values at the points are sometimes undefined.

The divergence of the gravitational (or Coulomb electrostatic) field was one of his first examples. Today the function (say, of one variable x), equal to zero everywhere except at $x = 0$ and having integral 1, is called “the δ-function of Dirac”. The divergence of the gravitational field is, up to the constant factor $-c$, the n-dimensional δ-function.

The general theory of such functions is today called “distribution theory” and “generalised function theory” (the names were given by L. Schwartz and S. Sobolev, respectively, who were the followers of Günter).

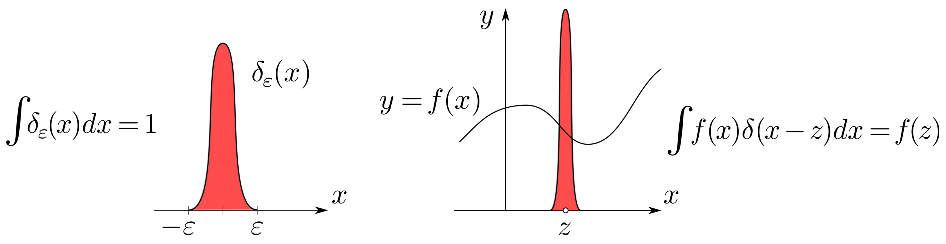


Figure 6.16: A δ-like function on a line and its integrals.

The easiest way to understand the formulas which contain such functions, is to replace the δ-function by an "approximating" positive smooth function δ_ε, depending on a small parameter ε in such a way that the function vanishes everywhere, outside an ε-neighbourhood of the origin, having however an integral equal to 1 (having, then, very large values in the smaller neighbourhood of the origin, Fig. 6.16).

A relation containing the δ-function means the limit of the corresponding relations, involving the approximating functions δ_ε for ε tending to zero.

Example. The integral

$$f(z) = \int f(x)\ \delta(x-z)\ dx$$

can be interpreted as the representation of any smooth function f (of variable x) in the form of a "continuous linear combination" (that is, of an integral) of the special functions $\delta(x-z)$ of the variable z (this is the usual δ-function, shifted from 0 to the point z). In the above combination, these basic functions $\delta(x-z)$ of argument x, have some numerical coefficients which are the values $f(x)$ of the function f at different points x.

Hence, the (shifted) δ-functions form a "basis" of the space of smooth or continuous functions – which is extremely important for the calculations of the linear and nonlinear reactions (called also "replies") of the systems under exterior perturbations. In the linear case, a typical example is the Poisson formula for the potential of a distributed mass or electric charge: one should integrate the "fundamental solutions", corresponding to the shifted δ-perturbations with suitable coefficients.

Günter created his theory for nonlinear hydrodynamical applications. He deduced from it some nontrivial existence and uniqueness theorems for the Navier–Stokes equations, especially in dimension 2, which were later extended to other functional spaces by

Ladyzhenskaya. He proposed to his student Sobolev to use his "functions of domains" to the study of linear problems of wave propagation in acoustics and seismology, studying hyperbolic differential equations. This applications of his theory to practical problems (by a student, member of the Communist party) were very important for the life of Günter, who had been just accused "to promote aristocratic science unneeded to the proletarians".

L. Schwartz told to Arnold that Sobolev made a serious mistake publishing his brilliant results in a dull provincial journal, using a sparse language, and that the main Schwartz's contribution had been to translate it to English. To the natural question on the sparse language and the dull provincial journal, Schwartz answered that the language was the French, and the journal was Comptes Rendus de l'Académie des Sciences, Paris.

Sobolev, however, objected the Schwartz version of the story, telling to Arnold that Schwartz had added also his own important contributions to the theory, especially in what concerned the Fourier representation of distributions.

Distribution theory is a large and important subject in itself. Here, we shall only mention the following geometric problems relating homogeneous functions, the δ-function and the Gaussian curvature.

Definition. A function f is said to be *homogeneous of degree* d, if (say, for all $\lambda > 0$) $f(\lambda x) = \lambda^d f(x)$.

Problem. 1) Is the δ-function homogeneous (and what is its degree, if it is)?

2) Same question for the δ-function in $\mathbb{R}^n$.

Solution. Calculating the integral of $\delta(2x)$, we find

$$\int \delta(2x)\ dx = \int \delta(y)\ d(y/2) = \frac{1}{2}\int \delta(y)\ dy = \frac{1}{2}\ ,$$

therefore $\delta(2x) = \frac{1}{2}\delta(x)$, and hence $\deg\delta = -1$.

In the case of $\mathbb{R}^n$ we may either repeat the same reasoning, or use the (obvious) representation $\delta(x_1, \ldots, x_n) = \delta(x_1)\delta(x_2)\ldots\delta(x_n)$, proving also that, in the space $\mathbb{R}^n$, $\deg\delta = -n$.

Problem. Calculate the second derivative of the function $|x|$ of the real variable x.

Solution. The function $|x|$ is homogeneous of degree 1. Thus its second derivative should be homogeneous of degree -1. For any $x \neq 0$ the second derivative of $|x|$ evidently vanishes. We conclude that $d^2|x|/dx^2 = C\delta(x)$.

To calculate the constant C, note that the integral of the second derivative provides the increment of the first one. Since the first derivative is equal to sign(x), it has increment $+2$ (say, between $x = -1$ and $x = 1$). Therefore, $C = 2$, and we obtain the important formula

$$\frac{d^2|x|}{dx^2} = 2\delta(x) \ .$$

Problem. Calculate the Laplace operator Δ of the function $1/r$ in Euclidean space $\mathbb{R}^3$, where r is the distance to the origin $r = \sqrt{x^2 + y^2 + z^2}$, in the standard Cartesian coordinates.

Solution. The function $1/r$ is spherically symmetric and homogeneous of degree -1. Hence, $\Delta(1/r)$ is a spherically symmetric homogeneous function of degree -3. The gradient of $1/r$ is the gravitational field (see p. 187)

$$\operatorname{grad} \frac{1}{|X|} = -\frac{X}{|X|^3} \ .$$

Now, by the definition of the Laplace operator, $\Delta = \operatorname{div} \operatorname{grad}$, we have

$$\Delta(1/|X|) = \operatorname{div} \left(-\frac{X}{|X|^3} \right) = -4\pi\delta(X)$$

(see the divergence calculation on p. 206).

Problem. Find the Gaussian curvature of the cone whose rays form the angle $\alpha = 30°$ with its axis.

Solution. Inscribing a sphere of radius 1 in the cone (Fig. 6.21), we construct a convex surface consisting of two parts. One of them is formed by the segments (of the generatrices) which join the vertex of the cone to the point touching the sphere; and the other part is the segment of the sphere opposite to the vertex of the cone. The height of this segment is $1 + \sin\alpha = 3/2$. The areas of the segments of the sphere are proportional to their heights. Hence the area of our segment is $4\pi(1 + \sin\alpha)/2 = 3\pi$.

Consequently, the integral of the Gaussian curvature of the smooth part of our convex surface is 3π. Since the total integral of the Gaussian curvature of a surface of genus 0 is 4π (see p. 376-377), the contribution of the singular point of the cone to the integral is π.

Outside the vertex the Gaussian curvature of the cone vanishes. In consequence, the Gaussian curvature of the cone is $2\pi(1 - \sin\alpha)\delta = \pi\delta$.

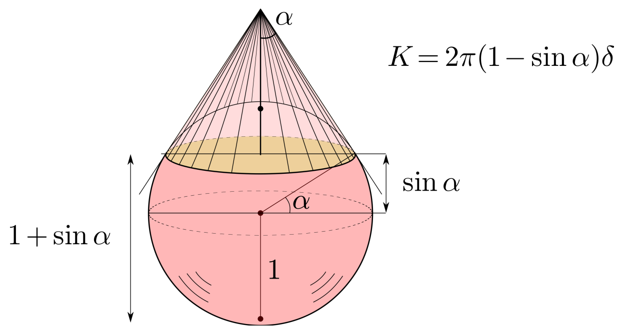


Figure 6.17: The calculation of the Gaussian curvature of a circular cone.

The curvature is diminishing when α approaches the angle $\pi/2$ (at which the cone becomes the tangent plane to the sphere), and the maximal curvature, $2\pi\delta$, corresponds to $\alpha = 0$ (when the cone approaches the cylinder, the singular point being infinitely far from the inscribed unit sphere).

Problem. Find the Gaussian curvature of the cube.

Solution. The Gaussian curvature of any surface consisting of parallel lines is 0. Thus the curvature vanishes along the faces and along the edges of the cube, outside the 8 vertices.

The δ-curvatures at the 8 vertices have equal coefficients c, since the curvature is invariant with respect to the isometries of the cube.

The integral of the Gaussian curvature along a convex closed surface is 4π (see Ch. 10, p. 376-377, where the Gauss-Bonnet formula is discussed). Thus we get $c = \pi/2$:

$$K = \sum_{i=1}^{8} \left(\frac{\pi}{2}\right) \delta_i \ ,$$

where δ_i denotes the δ-function concentrated at the i-th vertex.

6.8 Properties of the exterior derivative

There is a lot of differential forms whose exterior derivative vanishes identically. We have already seen many examples of them.

Example. For the differential of any smooth function of 2 variables

$$df = P\ dx + Q\ dy\ , \quad P = \partial f/\partial x\ , \quad Q = \partial f/\partial y\ ,$$

the exterior derivative 2-form is

$$d(df) = dP \wedge dx + dQ \wedge dy = \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx \wedge dy =$$
$$= \left(\frac{\partial^2 f}{\partial x \partial y} - \frac{\partial^2 f}{\partial y \partial x}\right) dx \wedge dy = 0 \ .$$

A similar calculation shows that $d(df) = 0$ for any smooth function of n variables.

These computations mask the deep geometric reason for which the second exterior derivative vanishes, which is the following theorem of elementary stereometry.

Theorem 5. *The oriented boundary of the oriented boundary of any convex polyhedron in Euclidean space vanishes:* $\partial(\partial P) = 0$.

Example. A triangle ABC on the plane is a 2-chain, oriented by the frame $(\overrightarrow{AB}, \overrightarrow{AC})$, whose boundary is the sum of the three oriented 1-chains AB, BC and CA, with coefficients 1 (Fig. 6.18).

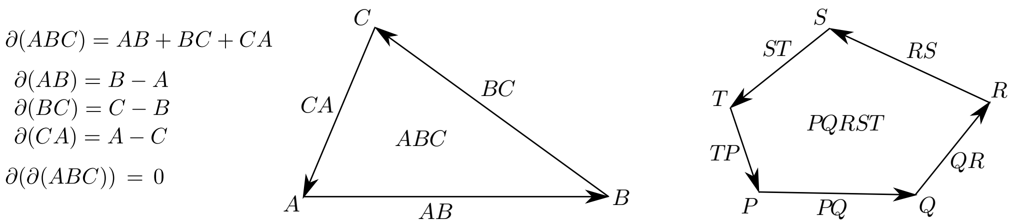


Figure 6.18: The boundary of the boundary vanishes for a convex polygon.

The boundary of each oriented segment is a 0-chain. Namely, it is the difference between the final point and the starting point of the oriented segment ($\partial(AB) = B - A$ and so on).

The boundary of the boundary is $\partial(\partial(ABC)) = \partial(AB) + \partial(BC) + \partial(CA)$. Each of the three vertices A, B and C enters twice in the resulting sum of the boundaries of the three segments: once as a final point and once as a starting point. Hence the total sum vanishes.

A similar reasoning holds for any convex polygon in the plane, as for the pentagon $PQRST$ of Fig. 6.18, oriented by the frame $(\overrightarrow{PQ}, \overrightarrow{PT})$.

Strangely enough, the general result, $\partial^2 = 0$, for the polyhedrons of any dimension n, follows from the result of this example.

Indeed, the boundary of the boundary of a convex n-dimensional polyhedron is, by definition, an $(n-2)$-dimensional chain, consisting of the $(n-2)$-dimensional faces of the original polyhedron with some integer coefficients.

To compute the coefficient of a face f of dimension $n-2$, consider the transverse section of the polyhedron P by the 2-dimensional orthocomplement V to the face f. It intersects P along a convex polygon whose sides are the intersections of V with ∂P, and whose vertices are the intersections of V with $\partial(\partial P)$ (including f).

To compute the coefficient we should compare the signs of the frames of these faces. But we shall avoid this job, choosing some frame in f and continuing it to frames in the faces of ∂P which contains f, and in P: to do it we add the exterior normals which orient the sections with the plane V of the faces (we are considering).

In this way we obtain the following natural conclusion: The boundary of the intersection of a face with V is the intersection of the boundary of this face with V (taking the orientations of the boundaries into account).

In consequence, the vanishing of the coefficient of a point of the second boundary for the polygonal section of P by V proves the vanishing of the coefficient of the corresponding face of the second boundary of P.

Essentially it proves the identity $\partial^2 P = 0$, but to be precise one should guarantee that the 2-dimensional section V intersects transversely (at isolated points) the $(n-2)$-dimensional faces of the chain $\partial^2 P$.

To achieve this, if it were not the case, it would be sufficient to replace the plane V by a slightly perturbed two-plane, and the argument of the convex polygon (Fig. 6.18) would be applied to the perturbed section.

Corollary. *The second exterior derivative of any smooth exterior differential form vanishes identically:* $d(d\omega) = 0$.

Proof. Integrate the second exterior derivative of an m-form ω along a convex polyhedron P of dimension $m+2$.

By the Stokes Theorem, this integral is equal to the integral of the first exterior derivative along the boundary chain of P,

$$\iiint\limits_{P} d(d\omega) = \iint\limits_{\partial P} d\omega \ .$$

Using the Stokes theorem once more (applying it, in case of doubt, separately to each face of the boundary polyhedron ∂P, and adding the results) we get the identity

$$\iint_{\partial P} d\omega = \int_{\partial(\partial P)} \omega \ .$$

The last integral vanishes, since the second boundary chain $\partial(\partial P)$ is zero. This proves that the original integral

$$\iiint_{P} d(d\omega) = 0$$

vanishes for all the convex polyhedrons P of dimension $m+2$.

Consequently, the $(m+2)$-form $d(d\omega)$ vanishes identically, and the theorem is proved: $d^2 = 0$. □

6.9 Informal Duality Chaines $\leftrightarrow$ Forms

Of course, the identity $d(d\omega) = 0$ may be also obtained by a (long) calculation of the coefficients of the $(m+2)$-form $d(d\omega)$. But our geometric reasoning provides more than this algebra.

Namely, the relation $d^2 = 0$ for the forms follows, by the linear algebra duality, from the dual relation $\partial^2 = 0$, taking into account the (informal) duality between the space of the chains and the space of the forms.

This duality reasoning is extremely powerful, and we shall use it many times to prove (or at least to guess) different results of the algebra of differential forms from the dual (and elementary) facts of the geometry of chains.

The strict mathematical foundations of these duality reasonings exist, but they need some rather long and detailed studies of the corresponding infinite dimensional spaces of forms and of chains with their topologies and norms, corresponding to the different smoothness classes of the objects that we wish to study.

In most cases, it is easier to use the duality as informal guide, proving directly the theorems that the duality approach suggests, with no references to the complicate general functional analysis. We shall see below some interesting examples of this strategy.

Remark. Our forms-chains duality reasoning is not formalised for the following reason: the vector spaces of the smooth differential forms and of the naive smooth chains considered above, are not genuinely dual vector spaces in the sense of linear algebra, since for some nonzero naive chains the integrals of all the smooth forms, along one of those chains, are zero.

For instance, an (orientation preserving) diffeomorphism of the parametrising polyhedron provides formally a new naive chain, while all the forms have the same integral along this new chain, as along the old one ($\int_{x=0}^{x=1} f(x)\ dx = \int_{y=0}^{y=1/2} f(2y)\ d(2y)$).

To reach the genuine algebraic duality, one has either to consider more (nonsmooth) forms, or to diminish the vector space of the chains. Both ways are realisable, but we shall not follow below their complicate formalisms.

However, it is possible to identify those naive chains which are reducible one to other by the (orientation preserving) diffeomorphisms of the parametrising polyhedrons. Such sophisticated chains (equivalence classes) are still convenient integration domains (of all the forms) and still do form vector spaces. Thus the word "chain" in our theorems and proofs may mean these sophisticated chains (instead of the naive chains, defined formally on p. 193) – the text is written in such terms that both understandings of the word "chain" lead to the same conclusions.

In other words, we shall prove that $2+3=3+2$ by counting both sums and finding that both are equal to 5, rather than by a reference to "the general axiom on the commutativity of the addition operation", which would be a shorter (but less convincing) proof.

Theorem 6. *The exterior derivative of the sum of two smooth differential m-forms (defined on the same manifold) is the sum of their exterior derivatives.*

Proof. Of course, it follows directly from the derivative definition. The duality approach provides the same result, since the integral of the sum is equal to the sum of the integrals of the summands. □

Theorem 7. *The exterior derivative $d(\alpha \wedge \beta)$ of the exterior product of two smooth differential forms, where α is a k-form and β is an ℓ-form on the same manifold, is provided by the exterior derivatives of the factors in the following way:*

$$d(\alpha \wedge \beta) = (d\alpha) \wedge \beta + (-1)^k\ \alpha \wedge (d\beta)\ . \tag{6}$$

Proof. We can consider the "Leibniz rule" as a computational corollary of the coordinate calculation formula for the monomials

$$d(f\ dx_1 \wedge \cdots \wedge dx_m) = (df) \wedge (dx_1 \wedge \cdots \wedge dx_m)\ .$$

Let the forms α and β be the monomials

$$\alpha = a(dx_1 \wedge \cdots \wedge dx_k) \ , \quad \beta = b(dy_1 \wedge \cdots \wedge dy_\ell) \ .$$

Then we calculate the monomial expressions

$$\alpha \wedge \beta = ab(dx_1 \wedge \cdots \wedge dx_k \wedge dy_1 \wedge \cdots \wedge dy_\ell) \ ,$$

$$d\alpha = (da) \wedge (dx_1 \wedge \cdots \wedge dx_k) \ , \quad d\beta = (db) \wedge (dy_1 \wedge \cdots \wedge dy_\ell) \ .$$

The first expression provides the binomial formula

$$d(\alpha \wedge \beta) = (a\ db + b\ da) \wedge (dx_1 \wedge \cdots \wedge dx_k \wedge dy_1 \wedge \cdots \wedge dy_\ell) \ .$$

The monomials $(d\alpha) \wedge \beta$ and $\alpha \wedge (d\beta)$ have the expressions

$$(d\alpha) \wedge \beta = b\ da \wedge (dx_1 \wedge \cdots \wedge dx_k \wedge dy_1 \wedge \cdots \wedge dy_\ell) \ ,$$

$$\alpha \wedge (d\beta) = a(dx_1 \wedge \cdots \wedge dx_k) \wedge (db) \wedge (dy_1 \wedge \cdots \wedge dy_\ell) \ .$$

To obtain formula (6), it remains to move the 1-form db in the last formula, by jumping over the k 1-forms dx_i. These k jumps provide the sign $(-1)^k$ in the expression (6).

Formula (6) is thus proved for any two monomials α and β. But since every differential form is a sum of monomials, the distributivity of the multiplication with respect to the addition implies that the bilinear relation (6) holds for any smooth differential k-form α and any smooth differential ℓ-form β. Theorem 7 is proved. □

Remark. The duality approach is also related to the "Leibniz formula" for the boundary of the Cartesian product of oriented polyhedra:

$$\partial(P \times Q) = ((\partial P) \times Q) \ + \ \left((-1)^{\dim P} P \times (\partial Q)\right) \ .$$

We shall use this formula later, but it is already useful to check the signs, say for $\dim P = \dim Q = 1$, $\dim P = 1, \dim Q = 2$, $\dim P = 2, \dim Q = 1$, $\dim P = \dim Q = 2$ (see Fig. 6.19).

When $\dim P = 1$, it is useful to remember that the preceding formula takes the form

$$\partial(P \times Q) + (P \times \partial Q) = (\partial P) \times Q \ .$$

For example, for $P = [0,1]$ it means

$$\partial([0,1] \times Q) + ([0,1] \times (\partial Q)) = (1 \times Q) - (0 \times Q) \ .$$

Here the left hand side contains both products with the plus signs, while the right hand side of the formula represents the "increment of state" of Q.

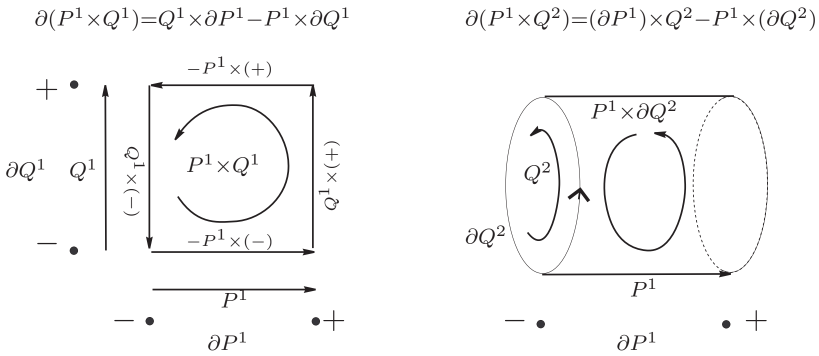


Figure 6.19: Oriented boundaries of direct products.

6.10 Symplectic and Contact Forms

Problem. Calculate the exterior derivative of the symplectic 2-form in $\mathbb{R}^{2n}$

$$\omega = dp_1 \wedge dq_1 + \cdots + dp_n \wedge dq_n \ .$$

Solution. Consider the "action 1-form" $\alpha = p_1 \ dq_1 + \cdots + p_n \ dq_n$. Then $\omega = d\alpha$, and hence $d\omega = d(d\alpha) = 0$.

Problem. Calculate the exterior derivatives of the exterior products of ω with itself.

Solution. The exterior derivatives of all the exterior products of a symplectic structure with itself vanish identically: $d(\omega \wedge \omega) = (d\omega) \wedge \omega + \omega \wedge (d\omega) = 0$, since $d\omega = 0$. Next,

$$d(\omega \wedge \omega \wedge \omega) = (d\omega) \wedge (\omega \wedge \omega) + \omega \wedge (d(\omega \wedge \omega)) = 0 \ ,$$

since $d\omega = 0$ and $d(\omega \wedge \omega) = 0$, and so on. The solution of this problem is known as "the Poincaré theorem on the integral invariants".

Problem. Find the exterior derivative of the 3-form $\alpha \wedge \omega$ (product of the action form α and of the symplectic structure $\omega = d\alpha$).

Solution. $d(\alpha \wedge \omega) = (d\alpha) \wedge \omega - \alpha \wedge (d\omega) = \omega \wedge \omega$, which is a nondegenerate 4-form in $\mathbb{R}^{2n}$, for $n > 1$ (see p. 182).

Problem. Consider a 1-form α and its exterior derivative ω, defined on a smooth manifold M of dimension $2n + 1$. Prove that the exterior product of degree $2n+1$ $\alpha \wedge \omega \wedge \cdots \wedge \omega$ is a non degenerate volume element if and only if the restriction of the 2-form ω to the hyperplane $\alpha = 0$ of the tangent space is a non degenerate 2-form.

Solution. Consider a parallelepiped of dimension $2n+1$ in the tangent space T_xM, formed by $2n$ vectors in the hyperplane $\alpha = 0$ and by one vector ξ on which $\alpha(\xi) \neq 0$.

The value of the product $\alpha \wedge \omega \wedge \cdots \wedge \omega$ on this parallelepiped consists only of one summand, $\alpha(\xi)c$, where c is the value of the product $\omega \wedge \cdots \wedge \omega$ on the $2n$ vectors of the plane $\alpha = 0$.

The non-degeneracy of the volume element means that $c \neq 0$.

If the 2-form ω on the $2n$-dimensional space is non-degenerate, then its n-th exterior power is the volume element in the $2n$-space. For instance, one can use the Darboux coordinates $(p_1, \ldots, q_n)$ in the $2n$-space, in which $\omega|_{\alpha=0}$ would have the Darboux form $(p_1 \wedge q_1) + \cdots + (p_n \wedge q_n)$ and to calculate explicitly the n-th power of this form, to see that it does not vanish (see page 182).

We say that the restriction of the 2-form ω to the hyperplane $\alpha_x = 0$ is degenerate, if there exists a non zero vector η (in the plane $\alpha_x = 0$), such that $\omega(\eta, \zeta) = 0$ for every vector ζ of the hyperplane $\alpha_x = 0$.

In this case $c = 0$, since to evaluate the product $\omega \wedge \cdots \wedge \omega$ at a basis, containing the kernel vector η, we would have factors of the form $\omega(\eta, \zeta)$, making the corresponding summand of the value of the product on our basis to vanish (and thus proving, that $c = 0$ whenever the 2-form ω is degenerated on the hyperplane $\alpha = 0$).

The above problem is basic for the so-called *contact geometry*: A 1-form α is called a *contact form* on an odd dimensional manifold, if at each point its exterior derivative $d\alpha$ is non-degenerate on the hyperplane of its zeroes. The field of these hyperplanes is called a *contact structure*.

The manifolds endowed with a contact structure are called *contact manifolds* and the study of their properties, maps, submanifolds, etc., is called *contact geometry*. Contact geometry is a very rich subject and has many relations with differential geometry, geometrical optics, wave front propagation, thermodynamics, control theory, etc. Chapter 16 is devoted to contact geometry together with its even-dimensional twin, symplectic geometry.

6.11 Closed forms, exact forms, Poincaré lemma

We already know several examples of differential k-forms whose exterior derivative vanish (including, for instance, all the exterior derivatives of the differential $(k-1)$-forms). Such forms have a name:

Definition. A differential k-form is called *closed*, if its exterior derivative vanishes.

This name was chosen, according to the (informal) duality, because a chain is closed if its boundary vanishes.

It is natural to ask whether there exist other "closed" k-forms?

The example of the 1-form

$$\omega = \frac{x\ dy - y\ dx}{x^2 + y^2}$$

on $\mathbb{R}^2 \setminus 0$ shows that there exist closed forms that are not exterior derivatives, that is, this closed 1-form is not the differential of a 0-form (a function).

Indeed, the integral of this form along the closed circle ($x = \cos t, y = \sin t$) does not vanish:

$$\oint_0^{2\pi} \cos t\ d(\sin t) - \sin t\ (d\cos t) = \oint_0^{2\pi} dt = 2\pi.$$

But the integral of the differential of a function would be the increment of that function, which is 0 for any closed path.

The exterior derivative of this 1-form is everywhere zero in $\mathbb{R}^2 \setminus 0$. This derivative 2-form is $(\mathrm{div}\ v)\ dx \wedge dy$, where

$$v = \left(-y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y} \right) r^{-2}\ .$$

This vector field v is directed along the circles $x^2 + y^2 = \mathrm{const}$, moving each of them with a constant angular velocity. Such motion preserves the areas, hence $(\mathrm{div}\ v) = 0$.

We can express the form ω in polar coordinates ($x = r\cos\varphi, y = r\sin\varphi$) as

$$\omega = d\varphi,$$

and in this sense ω is the exterior derivative of φ. However the "potential" φ is a multivalued function in $\mathbb{R}^2 \smallsetminus 0$, which cannot be considered there as a differential 0-form.

Exact Form. A form that is an exterior derivative is called *exact form*.

The existence of closed forms on a manifold M that are not exact is related to the topological properties of M.

Theorem 8 (Poincaré Lemma). *Every closed smooth differential k-form on Euclidean space $\mathbb{R}^n$ is the exterior derivative of a differential $(k-1)$-form.*

Proof. We start from the "dual" statement for chains:

Lemma. *Every chain σ^k in Euclidean space $\mathbb{R}^n$, whose boundary vanishes, is the boundary of some $(k+1)$-chain: if $\partial\sigma^k = 0$, then there exists a chain Σ^{k+1} for which $\sigma^k = \partial\Sigma^{k+1}$.*

Proof of the Lemma. We define the *cone* over a k-chain in $\mathbb{R}^n$ as the $(k+1)$-chain whose elements are obtained from the corresponding elements of the given k-chain by taking the union of the segments which join the points of those elements to the origin O of $\mathbb{R}^n$ (Fig. 6.20).

The orientation of an element of the cone is provided by the frame formed by the connecting vector ξ (starting at O), followed by the orienting frame of the initial element.

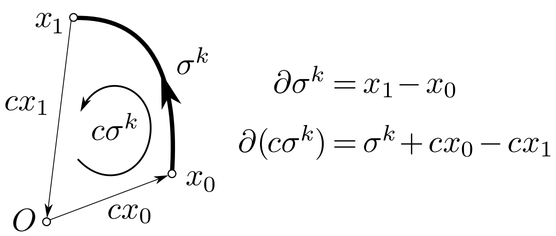


Figure 6.20: Boundary of a cone over a chain: $\partial c = 1 - c\partial$.

Denote by c the linear operation of the construction of the cone. Then the boundary of the resulting cone is (evidently) given by (Fig. 6.20):

$$\partial c(\sigma) = \sigma - c(\partial\sigma) \ .$$

It is convenient to write this relation in the easily memorisable form, without minus signs:

$$\partial c + c\partial = 1 \ .$$

Applying it to the chain σ of the lemma, we get $\partial(c\sigma) + c(\partial\sigma) = \sigma$. It follows that σ is represented as the boundary of the cone $c\sigma$, since $\partial\sigma$ is 0 by the condition of the Lemma, which is thus proved. □

Comment. One should remark that this easy proof is based on the generality of the notion of chains, which were not supposed to be smooth or even topological manifolds. The cone over a smooth submanifold is not smooth at all, but it is as well useful.

Co-cone Operation. Returning to the differential k-form ω, the duality arguments lead us to discover the following deep geometrical fact.

Lemma. *There exists a "co-cone" linear operation c^* sending the smooth differential k-forms in Euclidean space $\mathbb{R}^n$ to the $(k-1)$-forms, and verifying the condition*

$$dc^* + c^*d = 1 \ .$$

Of course, to guess it one has to start from the cone linear operator

$$c : \{k\text{-chains}\} \longrightarrow \{k+1\text{-chains}\} \ ,$$

and then to construct the dual linear operator c^*:

$$\{k\text{-chains}\}^* \longleftarrow \{k+1\text{-chains}\}^* \ .$$

In other words, the integral of the form $c^*\omega$ along any chain σ^k should be equal to the integral of the original $(k+1)$-form ω along the $(k+1)$-dimensional cone $c\sigma^k$.

In this way, we obtain the integrals of the form $c^*\omega$ along many chains, and then we reconstruct its values at k given tangent vectors $(\xi_1, \ldots, \xi_k)$ at the point x.

Namely, we construct a small chain $\sigma^k(\varepsilon)$, represented by the oriented parallelepiped with edges $(\varepsilon\xi_1, \ldots, \varepsilon\xi_k)$ at the point x. We take the integral of the unknown form $c^*\omega$ along this chain,

$$I(\varepsilon) = \int_{\sigma^k(\varepsilon)} c^*\omega = \iint\limits_{c\sigma^k(\varepsilon)} \omega \ ,$$

and then we obtain the value of the unknown form:

$$(c^*\omega)(\xi_1, \ldots, \xi_k) = \lim_{\varepsilon\to 0} \left(I(\varepsilon)/\varepsilon^k\right) \ .$$

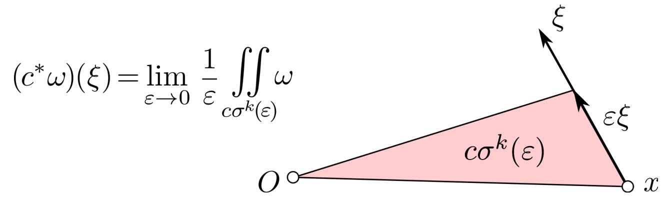


Figure 6.21: Proof of the Poincaré Lemma.

This limit can be written more explicitly as an integral along the segment $c(x)$ of the values of the original form ω on $k+1$ convenient vectors with directions $(x, \xi_1, \ldots, \xi_k)$, but the Poincaré theorem is already proved even if we are too lazy to write this integral formula explicitly. $\square$

6.12 Harmonic functions and their averages

Definition. A function in Euclidean space $\mathbb{R}^n$ is called *harmonic* if it verifies the Laplace equation

$$\Delta u = 0 \ , \quad \text{that is,} \quad \operatorname{div} \operatorname{grad} u = 0 \ . \tag{7}$$

PROBLEM. Find all the spherically-symmetric harmonic functions $u = f(r)$ on the complement to the origin in Euclidean space $\mathbb{R}^n$.

SOLUTION. The gradient at a point $x \in \mathbb{R}^n$ is $\operatorname{grad} u = (f'(r))x$, since the gradient vector at x should have the direction of the vector $x \in \mathbb{R}^n$, due to the spherical symmetry.

The flux of this vector field across the sphere $|x| = r$ is $f'(r)Cr^{n-1}$, where C is the $(n-1)$-dimensional volume of the sphere of radius 1.

If the function u were harmonic everywhere inside the ball $|x| \leqslant r$, then the flux across its boundary should vanish, since it is equal to the integral of the divergence of the gradient of u inside the ball (which vanishes, according to the Laplace equation (7)).

If the function u is harmonic everywhere, except the origin, the fluxes across the concentric spheres of different radii should be equal (since their difference is the integral of the divergence of the gradient of u along the annulus between the two concentric spheres, on which the divergence vanishes according to Laplace equation (7)).

Thus, $f'(r)r^{n-1} = \text{const}$, and hence, for $n \neq 2$, we get the functions

$$f(r) = c_0 + c_1 r^{2-n} ,$$

which are harmonic.

For $n = 2$ we find $f'(r) = \text{const}/r$, therefore $f(r) = c_0 + c_1 \ln r$.

Theorem 9. *All spherically symmetric harmonic functions in $\mathbb{R}^n \setminus 0$ are linear combinations of two special functions: $u = 1$ and $u = r^{2-n}$ if $n \neq 2$; or $u = 1$ and $u = \ln r$ if $n = 2$.*

Example. For $n = 3$ we get the special harmonic function $u = 1/r$. Its gradient vector field provides the universal gravitation law and the Coulomb electrostatic vector field in $\mathbb{R}^3 \setminus 0$,

$$\operatorname{grad}(1/r) = -x/|x|^3 .$$

Similarly, for other values of n we get the "gravitation in dimension n" $\operatorname{grad}(1/r^{2-n}) = cx/(|x|^n)$, which holds also for $n = 2$:

$$\operatorname{grad}(\ln r) = x/|x|^2 .$$

Remark. The above theorem shows that f is homogeneous of degree $2 - n$. This suggests that the transcendental function $\ln r$ should be considered as a "homogeneous function of degree 0": $r^0 \sim \ln r$.

A different (physical) way to understand this conclusion is to calculate the Taylor series of r^ε for small ε:

$$r^\varepsilon = e^{\varepsilon \ln r} = 1 + \varepsilon \ln r + O(\varepsilon^2) ,$$

making the 2-plane generated by the functions 1 and r^ε to converge, for $\varepsilon \to 0$, to the plane generated by the functions 1 and $\ln r$.

Problem. Compute the mean value of a harmonic function on Euclidean space $\mathbb{R}^n$ along an $(n-1)$-dimensional sphere.

Answer. *The mean value of a harmonic function along a sphere is equal to its value at the centre of this sphere.*

Solution. Observe that the mean value along the concentric spheres is independent on the radius R. The limiting case $R \to 0$ provides, evidently, the value at the centre.

To simplify the formula, we choose the origin of Euclidean space to be the centre of our spheres.

In order to compare the mean values along two infinitesimally close concentric spheres, remark that the contributions of a small neighbourhood $d\omega$ of the point ω (of the standard sphere of radius 1) to the integrals of u along the spheres of radii $r + dr$ and r are

$$u((r + dr)\omega)\ (r + dr)^{n-1}\ d\omega$$

and

$$u(r\omega)\ r^{n-1}\ d\omega,$$

where $d\omega$ means the $(n-1)$-dimensional volume element.

Hence, the contributions to the mean values are $u((r + dr)\omega)\ d\omega/c$ and $u(r\omega)\ d\omega/c$, where the constant c is the volume of the sphere of radius 1.

Replacing $u((r + dr)\omega)$ by its Taylor expansion

$$u(r\omega) + \left(\frac{\partial u}{\partial r}\ \omega\right)\ dr + o(|dr|)\ ,$$

we deduce that the difference of the contributions to the mean values along the two neighbouring concentric spheres is given by the expression

$$\big(u((r + dr)\omega) - u(r\omega)\big)\ d\omega/c = \big((\text{grad}\ \ u, \omega)\ dr\big)\ d\omega/c + o(|dr|)\ .$$

Thus, the integral of this contribution of the sphere element $d\omega$, along the sphere of radius 1 of the vectors ω, is the flux of the vector field grad u across our sphere $r\{\omega\}$ of radius r, multiplied by the small constant dr/c, plus an $o(dr)$ correction.

The flux of the vector field grad u across our sphere $r\{\omega\}$ is equal to the integral of the divergence of this field along the inside ball. It vanishes, since u is a harmonic function: div grad $u \equiv 0$ according to the Laplace equation (7).

We have thus proved that the mean value is independent of the radius, proving then the answer to the above problem. This answer is called "the means theorem" of the theory of harmonic functions.

A second proof of "the means theorem"

We start by replacing the original harmonic function u by the functions obtained from it, using the rotations $g \in SO(n)$ of the Euclidean space around the centre O of our sphere. The rotated functions, $\widetilde{u}(x) = u(gx)$, are also harmonic. The mean values of the initial function and of the rotated function along the same sphere $|x| = r$ are equal.

Any linear combination of harmonic functions is a harmonic function, since the Laplace equation (7) is linear. Consequently, the average function of the rotated harmonic functions $\widetilde{u}$ (averaging them along SO(n)) is a harmonic function on Euclidean space $\mathbb{R}^n$.

The value of this averaged function at every point x of the sphere $|x| = r$ is equal to the mean value of u along this sphere (the rotations lead to x all the other points in turn).

At the origin all the rotated functions $\widetilde{u}$ have the same value $u(0)$, and hence the averaged harmonic function has there the same value.

Thus, the averaged harmonic function is constant along the boundary $|x| = r$ of the ball $|x| \leqslant r$. This constant is equal to the mean value of the initial function u

To prove that such a function is a constant and has the same value everywhere inside the ball, we shall use the following statement (proved below):

Maximum principle. *The maximum and the minimum of a harmonic function cannot be reached at an interior point of the domain where it is harmonic.*

Thus the maximal and minimal values of our function are equal, since (by the maximal principle) they are attained on the boundary $|x| \leqslant r$, on which the function is constant. In consequence, the function cannot reach neither larger nor smaller values inside, and hence it is constant everywhere.

Thus the value $u(0)$ is equal to the mean value of our harmonic function u along the sphere $|x| = r$.

This averaging geometrical reasoning provides a proof of the theorem on the mean values, which is astonishingly different from the preceding one.

Even more astonishing is the fact that neither the first, nor the second geometrical proof of the mean values theorem for the harmonic functions are discussed in the PDE textbooks.

Problem. Calculate the mean value of the function $1/|x|$ on Euclidean space $\mathbb{R}^3$ along the sphere of radius R centred at an arbitrary point $X \in \mathbb{R}^3$, $|X| \neq R$.

Solution. If $|X| > R$, the function $1/|x|$ is harmonic inside the ball $|x-X| \leqslant R$, and hence the mean value is equal to the value $1/|X|$ of the function at the centre of the sphere.

If $|X| \leqslant R$, the previous theory cannot be applied, since the singular point $x = 0$ belongs to the ball bounded by the sphere.

Denote the mean value of the function $1/|x|$ along the sphere of radius R with centre X by $w(X)$. This function w is harmonic in the domain $|X| \neq R$; since it is a "linear combination" of the harmonic shifts of the function $1/|x|$ (moving $x = 0$ to the points of the sphere of radius R centred at X, we get a harmonic function at X).

Due to the spherical symmetry of $1/|x|$, the function w is spherically symmetric: $w(X) = f(|X|)$. We have the answer $f(X) = c_1 + c_2/|X|$ (from the problem on p. 221) and it remains only to calculate the values of the coefficients c_1 and c_2, which are constant for $|X| \leqslant R$, and which, for $|X| > R$, have the constant values 0 and 1, as we already know.

It suffices to calculate $f(0)$.

If the centre if the sphere is $X = 0$, the function $1/|x|$ takes the constant value $1/R$ on the sphere of radius R, and its mean value is also $1/R$.

In consequence, $f(0) = 1/R$, and we obtain $c_1 = 1/R$, $c_2 = 0$ (the function w is harmonic at $X = 0$).

The final answer for the mean value w of the function $1/|x|$ on Euclidean space $\mathbb{R}^3$ along the sphere $\{x : |x - X| \leqslant R\}$ is:

$$w(X) = \begin{cases} \frac{1}{|X|} & \text{if } |X| > R\,, \\ \frac{1}{R} & \text{if } |X| < R\,. \end{cases}$$

Problem. Find the mean value w of the function $\ln r$ on Euclidean plane $\mathbb{R}^2$ along the circle $\{x : |x - X| = R\}$.

Answer.

$$w(X) = \begin{cases} \ln|X| & \text{if } |X| > R\,, \\ \ln R & \text{if } |X| < R\,. \end{cases}$$

It is strange that these geometric averages miss in most textbooks on PDE theory.

Proof of the maximum principle.

Lemma. *Let u be a smooth non-constant function vanishing at the boundary of a bounded domain G of Euclidean space $\mathbb{R}^n$. Then its second differential is either a positive definite or a negative definite quadratic form at some interior point of G.*

Proof. Let $\max u = M > 0$. Choose a linear function $\ell = \sum a_i x_i$ whose coefficients are sufficiently small, then on the boundary ∂G of the domain G the function $u + \ell$ is everywhere smaller than $M/2$ and $|\ell| \leqslant M/2$ everywhere in G. Thus the maximum of the function $u+\ell$ in G is attained at some interior point.

The second differential of $u + \ell$ at this point is nondegenerate, provided that the vector with components $-a_i$ is not a critical value of the map $\operatorname{grad} u$. According to Sard's Lemma, almost every vector is a noncritical value, and we can choose the small vector a in such a way that the second differential of u (and of $u + \ell$) at the maximum point is non-degenerate.

This non-degenerate form should be negative definite, since the critical point is a maximum.

Similarly, if $\min u = -M < 0$, then there is an interior minimum point of the function $u+\ell$, and the second differential of the function u is a positive definite quadratic form. The Lemma is proved. □

Corollary. *A harmonic function equal to a constant along the boundary of a closed domain, is equal to the same constant everywhere inside this domain.*

Proof. Otherwise, by the Lemma, there would exist an interior point at which the second differential is either a positive definite or a negative definite quadratic form, which is impossible since the trace of the second differential quadratic form of a harmonic function vanishes everywhere. □

Chapter 7

The fisherman (Lie) derivative

The main geometric derivative of a smooth differential form, of a smooth vector field or of any smooth tensor field on a manifold is the derivative along a given vector field v. The value of that derivative is, respectively, a differential form, a vector field or a tensor-field of the same kind as the original field which is derivated.

7.1 Geometric derivation and flows of vector fields

The difficulty of the usual derivation (like the partial derivative) is its non geometric definition. Indeed, say, the partial derivative $\partial f/\partial x$ is not well defined by the two functions f and x mentioned in this (disastrous) Leibniz notation: One should mention the whole coordinate system (say, the second variable y in the case of a function f of two variables: $\partial f/\partial x|_{y=\text{const}}$).

Example. (Fig. 7.1) Using the coordinate system $(x, \widetilde{y})$, where $\widetilde{y} = x + y$, we have

$$\frac{\partial f}{\partial x}|_{\widetilde{y}=\text{const}} = 0$$

for the same function on the plane $f = x + y$ for which

$$\frac{\partial f}{\partial x}|_{y=\text{const}} = 1.$$

This example shows that the usual notation $\partial f/\partial x$ is ambiguous, and to define the derivative one should mention the vector along which we wish derivate, rather than the coordinate x.

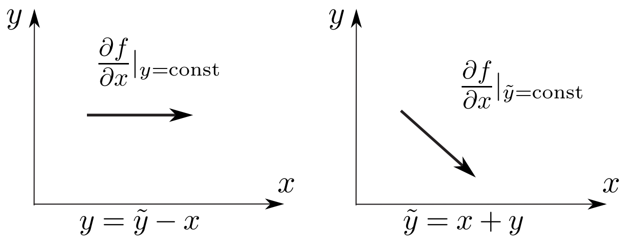


Figure 7.1: The ambiguity of the partial derivative $\partial f/\partial x$.

There are, however, a lot of partial derivatives and coordinate systems in physics, especially in thermodynamics. Of course, partial derivatives are also used throughout this book. The following paradoxical example is due to the greatest American physicist Gibbs, who invented it studying the partial derivatives of some thermodynamic potentials.

Suppose that a smooth function F of three variables (x, y, z) verifies the implicit function theorem, and that the surface M defined by the equation $F(x, y, z) = 0$ (in a neighbourhood of some point) is a graph of any of the three functions of two variables,

$$z = f(x, y), \quad x = g(y, z), \quad y = h(z, x).$$

Gibbs's problem. Calculate the product of the three partial derivatives

$$\left(\frac{\partial z}{\partial x}\right)\left(\frac{\partial x}{\partial y}\right)\left(\frac{\partial y}{\partial z}\right).$$

ANSWER. The bad notations suggest the answer 1, since, say, ∂z is once upstairs and once downstairs. The paradoxical answer of Gibbs is

$$\left(\frac{\partial f}{\partial x}\right)|_{y=\text{const}}\left(\frac{\partial g}{\partial y}\right)|_{z=\text{const}}\left(\frac{\partial h}{\partial z}\right)|_{x=\text{const}} = -1.$$

Pushkin mentioned "i genii, paradoxov drug" in a poem on the future of the sciences: "and the genial persons, friends of the paradoxes". Ya. B. Zeldovich, one of the fathers of the Russian H-bomb, took the pseudonym "Paradoxov", suggesting that it is the name of a friend of a genial person, according to Pushkin, and considering himself to be a friend of A. D. Sakharov.

The Gibbs paradox is only a rather easy objection to the derivation. A worse problem comes from the will to consider the "increment" of a tensor field*, which should measure the difference of the tensors at two neighbouring, but different, points x and x' of the manifold M where the tensor field is defined.

These two tensors belong to different vector spaces (say, $\Lambda^m(T_x^*M)$ and $\Lambda^m(T_{x'}^*M)$ for the case of the differential m-forms).

The identification of these two vector spaces (as well as the identification of the tangent spaces T_xM and $T_{x'}M$) is a non geometric construction. One usually uses the coordinate systems to identify both spaces to $\mathbb{R}^n$, where $n = \dim M$. But this identification would be different for a different coordinate system, and the resulting derivation would also depend on the choice of the coordinate system, having then no geometric meaning and even having no definition on a global manifold (like on a sphere), where there is no distinguished coordinate system.

One of the ways to overcome this crucial difficulty is to define some particular identification of the tangent spaces at two different points, called "connection of the tangent bundle". These connections are themselves very interesting objects of great usefulness in mathematical physics, and we shall say something on their theories later.

Nevertheless there is an intrinsic derivation of the tensor-fields on a smooth manifold M along a vector field v, which is independent of any connection or coordinate system – it is called the *Lie derivative* and is usually denoted L_v (for the mathematician Sophus Lie who used a lot these fisherman derivatives, known and used earlier by the best mathematicians and physicists). To introduce it we need the flows of vector fields (see also p. 38).

Phase Flow. Each vector field determines a *flow*: a one-parameter group of diffeomorphisms $g^t : M \to M,\ t \in \mathbb{R}$, satisfying $g^{s+t} = g^s g^t$.

The last identity, called the *group property* of the flow, means that the evolution law of the points in M does not depend on the time moment when we start to observe it (Fig. 7.2). The relation

$$g^{s+t}(x_0) = g^t(g^s(x_0))$$

means that the point $g^s(x_0)$ which started at x_0 in the moment 0 (in the right hand side), arrives after a time interval t at the same place in M where

---

*In an informal sense, a tensor is a generalised "linear object" that extends the notion of scalar, vector (multi-) linear functions, etc. For example, a Riemannian metric on a manifold (see footnote of p. 168) is a tensor field (some times called metric tensor (field)). A differential m-form is also a tensor field.

the point x_0 is sent by the evolution during a time interval $t+s$ (which is the left hand side).

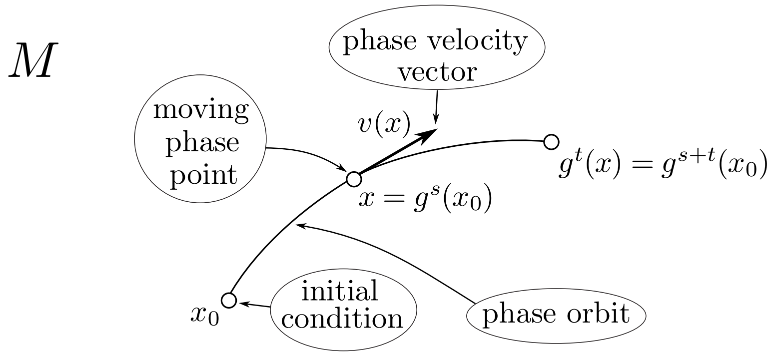


Figure 7.2: The flow $\{g^t\}$ determines its vector field $v(x) = \frac{d}{dt}|_{t=0} g^t(x)$ on M.

Thus the velocity of the motion at a given point x of M does not depend on the time moment:

$$\left(\frac{d}{dt}|_{t=0} g^t\right)(x) = \left(\frac{d}{dt}|_{t=0} g^{t+s}\right)(x_0) := v(x),$$

where $x = g^s(x_0)$.

Notation. Since the derivative of a function f along a vector field v with components $(a(x,y), b(x,y))$ is given by $v \cdot f = a\,\partial f/\partial x + b\,\partial f/\partial y$, one often considers and denotes vector fields as "derivation operators":

$$v = a\frac{\partial}{\partial x} + b\frac{\partial}{\partial y}.$$

Problem. Find the vector field of the one-parameter group of rotations defined by the matrices

$$g^t = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}$$

on Euclidean plane (equipped with orthonormal coordinates x and y).

Solution. Since $g^t \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\cos t + y\sin t \\ -x\sin t + y\cos t \end{pmatrix}$, we have $v\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ -x \end{pmatrix}$, that is,

$$v = y\frac{\partial}{\partial x} - x\frac{\partial}{\partial y}.$$

Problem. Let the phase space be the real line $M = \mathbb{R}$. Find the (phase) flow of the vector field of velocities $v(x) = x^2 d/dx$.

Solution. The motion $\varphi(t) = g^t(x_0)$ defined by the flow g must verify the differential equation which determines the phase velocity field v,

$$\frac{d}{dt}\varphi(t) = v(\varphi(t)), \tag{1}$$

with initial condition $\varphi(0) = x_0$.

In our case we get the following set of equalities

$$\frac{d\varphi}{dt} = \varphi^2, \quad \frac{d\varphi}{\varphi^2} = dt, \quad \int_{x_0}^{x} \frac{d\varphi}{\varphi^2} = t, \quad \frac{1}{x_0} - \frac{1}{x} = t, \quad x = \frac{1}{\frac{1}{x_0} - t},$$

obtaining the one-parameter family of maps (Fig. 7.3):

$$g^t(x_0) = \frac{x_0}{1 - tx_0}\,.$$

This family verifies the group property $g^{s+t} = g^s g^t$ and provides the velocity field $v(x) = x^2\partial/\partial x$, but is not a phase flow (i.e., it is not a one-parameter group of transformations) because for each $t \neq 0$ the map g^t is not a transformation of the line $M = \mathbb{R}$ (being not defined at the point $x_*(t) = 1/t$).

In consequence, the phase velocity vector field $v(x) = x^2\partial/\partial x$ on the real line ($\{x \in \mathbb{R}\}$) has no phase flow.

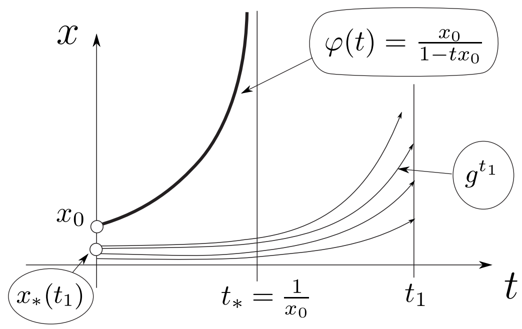


Figure 7.3: A smooth vector field on the real line with no phase flow.

Remark. The theory of ordinary differential equations guarantees the existence, uniqueness and smoothness (with respect to the initial conditions) of the solutions of a differential equation (defined by a smooth vector field) only for a small time interval that depends on the initial conditions (in Fig. 7.3 our solution $\varphi(t)$ is defined and smooth for $t < 1/x_0$).

The difficulty, in this case, is due to the non compactness of the phase space $M = \mathbb{R}$. On a compact manifold every smooth vector field has a phase flow which is defined for all t by the solutions $\varphi(t) = g^t(x)$ of the differential equation (1) with initial conditions $\varphi(t) = g^t(x_0)$ for $\varphi(0) = x_0$.

Problem. Prove that the vector field $v(x) = x^2\partial/\partial x$ on the affine line $\mathbb{R} = \{x\}$ is the restriction to the affine part of the projective line $\mathbb{R}\mathrm{P}^1$ of a smooth vector field there.

Solution. In fact, our phase flow $\{g^t\}$ consists of diffeomorphisms of the projective line $\mathbb{R}\mathrm{P}^1 \approx \mathbb{S}^1$ (rather than of diffeomorphisms of its affine part), and the "singularity" at $t = 1/x_0$ is simply the indication of the fact that $g^{t_*(x_0)}x_0 = \infty \in \mathbb{R}\mathrm{P}^1$.

Problem. Find the behaviour of the vector field $v(x) = x^2\partial/\partial x$ near the point $x = \infty$ of the projective line $\mathbb{R}\mathrm{P}^1$ with the affine coordinate x.

Solution. The affine coordinate $y = 1/x$ (smooth in a neighbourhood of the point $x = \infty$) has the derivative

$$\frac{dy}{dt} = -\frac{dx/dt}{x^2} = -\frac{x^2}{x^2} = -1.$$

Thus, near $y = 0$, the vector field $v(x) = x^2\partial/\partial x$ has the form $V = -1\cdot\partial/\partial y$. Hence the point $(x = \infty, y = 0)$ is a nonsingular point because this regular field V does not vanish.

Problem. Find the vector fields $x^n\partial/\partial x$ which are the restrictions to the affine line $\mathbb{R} = \{x\}$ of some smooth vector field on the projective line.

Answer. Only $n = 0, 1$ and 2.

Solution. For the vector fields on the line $v(x) = x^p\dfrac{\partial}{\partial x}$ with $p \in \mathbb{N}$ we use the coordinate $y = 1/x$ in the neighbourhood of $x = \infty$ and write this field as $v(y) = b(y)\dfrac{\partial}{\partial y}$. The function $b = b(y)$, obtained by derivating y along v,

$$b(y) = L_v y = x^p\frac{\partial}{\partial x}\left(\frac{1}{x}\right) = -x^p x^{-2} = -y^{-p}y^2 = -y^{2-p}\,,$$

is smooth if $p \leqslant 2$. The smooth fields have the form $v = (c_0 + c_1 x + c_2 x^2)\partial/\partial x$.

7.2 Fisherman Lie Derivative

On a non-compact manifold M a smooth velocity field v does not generate, in general, a phase flow. However, *locally* the flow $\{g^t\}$ is well defined: For $|\, t \,| < \varepsilon$ there exists diffeomorphisms of some neighbourhood U of the initial point x_0 of M,

$$g^t : U \to M, \quad \frac{d}{dt} g^t(x) = v(g^t(x)), \quad g^0(x) = x.$$

Such a *local phase flow* will suffice for our (following) definition of the derivative along a vector field, and we shall some times call it simply "the phase flow", meaning "the local phase flow".

The diffeomorphisms act on the tensor-fields, transporting them to new points by the derivative maps of the tangent and cotangent spaces (Fig. 7.4):

$$g^t : M \to M, \quad g^t_* : T_x M \to T_{g^t(x)} M, \quad g^{t^*} : T^*_{g^t(x)} M \to T^*_x M\,.$$

These linear isomorphisms produce similar isomorphic maps of the tensor products, symmetric tensor products and exterior tensor products of these vector spaces (discussed below, pp. 266-268).

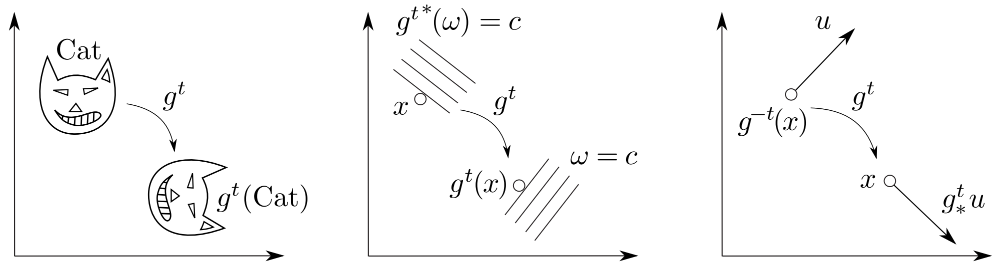


Figure 7.4: The phase flow transport of a 1-form ω and of a vector field u.

For instance, the exterior differential k-forms at the point $g^t x$ are transported by g^{t^*} to the point x, and the vectors at the point $g^{-t}x$ are transported by g^t_* to the point x.

Definition. The *Lie derivative* $L_v\omega$ of a smooth differential m-form ω at a point x of a smooth manifold M, derivated along a smooth vector field v on M, is the derivative with respect to the time of the time-dependent m-form at the point x, transported from the point $g^t(x)$ to x by the flow g^t of v:

$$L_v\omega = \frac{d}{dt}\Big|_{t=0}\; g^{t*}(\omega).$$

The main point of this definition is the fact that we differentiate with respect to time the variable vector $g^{t*}(\omega)$ *of a time-independent vector space*, $\Lambda^m(T_x^*M)$.

The flow $\{g^t\}$ transports the tensors near the place x where the fisherman waits for them, differentiating *in time* the passing tensors. He never compares vectors of different (x-dependent) vector spaces – Fig. 7.5.

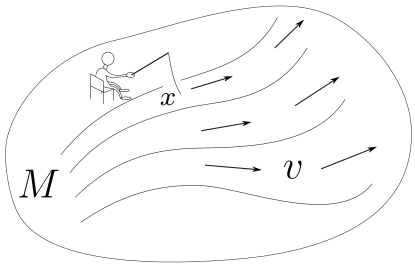


Figure 7.5: The fisherman differentiation in time.

Remark. For the functions (that is, for the 0-forms) the Lie derivative at x depends only on the value of the vector field at x, being then a "derivative along a vector": it is the usual "directional derivative of f along the vector v", $df(v)$ also denoted by $\partial_v f$. For the general case ($m \geqslant 1$) we are *not* defining any "derivative along a vector"; we shall explicitly see that the Lie derivative depends on the continuation of the given vector v at the point x to a vector field on its neighbourhood in M.

The Lie derivatives of the vector fields and of other tensor fields are defined similarly, transporting the tensors to x by the actions of g^{t*} and of g_*^t on the tensor products. An important example is the derivative of a vector field along another vector field (see p. 241).

7.3 Homotopy Formula

To calculate explicitly the Lie derivative, we shall prove the extremely important *Cartan homotopy formula*

$$L = id + di\,.$$

To explain it, we need to spell out the notations.

Let ω be a k-form on a vector space V and v be a vector.

Interior derivation. The $(k-1)$-form $i_v\omega$ is defined by the substitution of the vector v as the first argument of the function ω of k vectors:

$$i_v\omega(\xi_1,\dots,\xi_{k-1}) := \omega(v,\xi_1\dots,\xi_{k-1}).$$

In physics, the interior derivation i_v is mostly called "contraction with v".

Examples. 1. For $k=1$ the number $i_v\omega=\omega(v)\in\mathbb{R}$ is "a 0-form".
2. For $k=2$ and the area form $\omega = dx\wedge dy$, in oriented Euclidean plane $\mathbb{R}^2$ with standard Cartesian coordinates (x,y), we get the 1-form

$$i_{(P\,\partial/\partial x + Q\,\partial/\partial y)}(dx\wedge dy) = Pdy - Qdx$$

(whose integral is the flux of the vector field $v = P\ \partial/\partial x + Q\ \partial/\partial y$).
3. In Euclidean 3-space with the standard coordinates (x,y,z), the volume form $\tau = dx\wedge dy\wedge dz$ and a vector field $v = P\,\partial/\partial x + Q\,\partial/\partial y + R\,\partial/\partial z$ we get

$$i_v\tau = P\ dy\wedge dz + Q\ dz\wedge dx + R\ dx\wedge dy,$$

(whose integral along an oriented surface is the flux of the vector field v across that surface – the surface is cooriented in order that its orientation and coorientation determine a positive orientation of the 3-space).

Obvious properties. The operation $i_v : \Lambda^k \to \Lambda^{k-1}$ satisfies:

$$i_{cv} = ci_v\ \ (c\in\mathbb{R}),\quad i_{v+w} = i_v + i_w,\quad i_v i_w = -i_w i_v,\quad i_v(\alpha+\beta) = i_v\alpha + i_v\beta,$$

and $$i_v(\alpha\wedge\beta) = (i_v\alpha)\wedge\beta + (-1)^{\deg\alpha}\alpha\wedge(i_v\beta).$$

Theorem (Homotopy Formula). *The Lie* (fisherman) *derivation L_v along a smooth vector field v is related to the exterior derivation d and to the interior derivation i by the identity*

$$L_v\omega = i_v(d\omega) + d(i_v\omega)\,. \tag{2}$$

Remark. If ω is a k-form, $d\omega$ is a $(k+1)$-form, and hence $i_v(d\omega)$ is a k-form. Similarly, since $i_v\omega$ is a $(k-1)$-form, $d(i_v\omega)$ is a k-form. Thus the above sum is a k-form, as it ought to be for the Lie derivative. Let us see two examples.

Lie derivative of functions. For $k = 0$, ω being a smooth function, the homotopy formula provides the expression

$$L_v\omega = i_v(d\omega).$$

The second term, $d(i_v\omega)$, is zero, since there is no place to insert v in the 0-form ω. In this case we get simply the differential of ω applied to v

$$L_v\omega = d\omega(v) = \sum_{i=1}^{n} v_i \frac{\partial\,\omega}{\partial x_i}\,, \tag{3}$$

which is the usual directional derivative of the function ω along the vector $v(x) \in T_xM$. We see that this value, at the point x, does not depend on the continuation of the vector $v(x)$ to the field v.

Oriented volume and divergence zero. In oriented Euclidean plane with standard Cartesian coordinates (x, y), we write the homotopy formula in order to derivate the volume element $dx \wedge dy$ along the smooth vector field $v = P\,\partial/\partial x + Q\,\partial/\partial y$:

$$L_v(dx \wedge dy) = i_v d(dx \wedge dy) + d(i_v(dx \wedge dy)).$$

The first term vanishes, since $dx \wedge dy = d(x \wedge dy)$ and $d(d(x \wedge dy)) = d^2(\cdot) = 0$ (see p.212). The second term provides the answer (see p. 235):

$$d(i_v(dx\wedge dy)) = d(Pdy - Qdx) = (\partial P/\partial x + \partial Q/\partial y)(dx\wedge dy) = (\operatorname{div} v)(dx\wedge dy).$$

In consequence, the flow of a vector field v preserves the volume element $\tau = dx_1 \wedge \ldots \wedge dx_n$ of Euclidean space $\mathbb{R}^n$ (with Cartesian coordinates x_k) if and only if its divergence vanishes identically: $\operatorname{div} v \equiv 0$.

Of course, we have calculated $L_v\tau$ for the case $n = 2$, but the calculations are similar in the general case.

7.3.1 Proof of the Homotopy Formula

Of course, one may calculate the Taylor series of the coefficients of the transported form in a chosen coordinate system, and then (making no mistake in the notations of many indices) to compare the time derivative with the right hand side expression.

We prefer, however, a geometric proof in which the homotopy identity (2) is interpreted as a relation which is dual to an obvious geometric fact.

Homotopy operator. Given a smooth vector field v on a manifold M, the *homotopy operator* H transforms any oriented k-chain $\sigma : I^k \to M$ to the $(1+k)$-chain $H\sigma : I \times I^k \to M$ given by $(H\sigma)(s,t) = g^s\sigma(t)$, where the cube $\{s,t\}$ is oriented by the order $(s, t_1, \ldots, t_k)$ of the coordinates – Fig. 7.6.

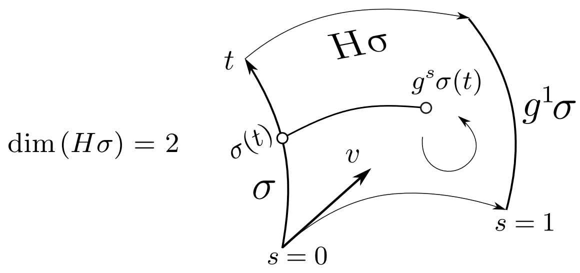


Figure 7.6: The homotopy $H\sigma$ of the 1-chain σ.

(To simplify notations, we have supposed that the flow $\{g^s\}$ is defined on σ for $0 \leqslant s \leqslant 1$. In the case of a local flow, the notations should be slightly different, but the geometric content is the same.)

Proof of Homotopy Formula. The core of the proof is the description of the boundary of the contructed chain $H\sigma$.

The "cylinder" $H\sigma$ has at least the two boundary parts corresponding to $s = 1$ and $s = 0$, even if the initial k-chain σ is closed ($\partial\sigma = 0$).

If the k-chain σ is not closed ($\partial\sigma \neq 0$), then the homotopy operator transforms the boundary $\partial\sigma$ into the k-chain $H(\partial\sigma)$, which is the remaining part of the boundary $\partial(H\sigma)$ of the $(k+1)$-chain $H\sigma$.

To control the signs, we use the obvious property of the boundary of a product: $\partial(I \times A) = ((\partial I) \times A) - (I \times (\partial A))$, that is,

$$\partial(I \times A) + (I \times (\partial A)) = (\partial I) \times A$$

(see p. 215 for a similar formula in the more general case of any two factors).

Applying this formula to the product $I \times I^k$, we get the following boundary relation of the chains (Fig. 7.7):

$$\partial(H\sigma) + H(\partial\sigma) = g^1\sigma - \sigma. \tag{4}$$

This geometric identity is the main content of the homotopy formula.

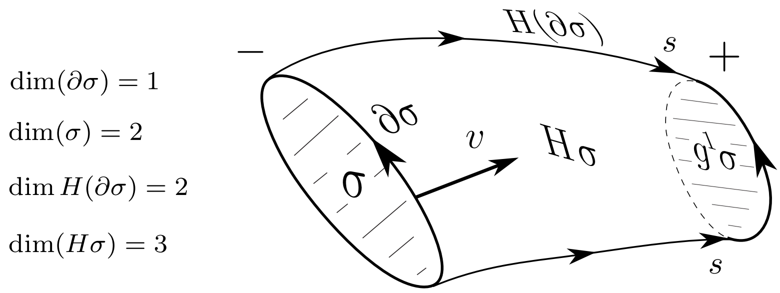


Figure 7.7: The boundary of the homotopy, $\partial(H\sigma) = -H(\partial\sigma) + g^1\sigma - \sigma$.

Integrating the k-form ω along the k-chains of the left hand side of (4), Stokes formula (of p.205) provides the equality

$$\int_{\partial(H\sigma)} \omega + \int_{H(\partial\sigma)} \omega = \int_{H\sigma} d\omega + \int_{H(\partial\sigma)} \omega \,.$$

In general, when one replaces the vector field v by the small field εv, the integral of any form α along HA is given by the expression $\varepsilon \int_A i_v\alpha + o(\varepsilon)$. In our case, the integral of the left hand side of (4) is (by Stokes formula)

$$\int_{H\sigma} d\omega + \int_{H(\partial\sigma)} \omega = \varepsilon \left(\int_\sigma i_v(d\omega) + \int_\sigma d(i_v\omega) \right) + o(\varepsilon) \,.$$

For the integral of the difference of the right hand side of (4), corresponding to the time 1 for the flow of the field εv (or equivalently to the time ε for the flow of the field v), we find

$$\int_{g^1\sigma} \omega - \int_\sigma \omega = \int_\sigma ({g^1}^* \omega - \omega) = \varepsilon \int_\sigma L_v\omega + o(\varepsilon).$$

Comparing the above integral of the left hand side of (4) with this expression for the right hand side, we conclude that the k-forms

$$i_v(d\omega) + d(i_v\omega) \quad \text{and} \quad L_v\omega$$

have equal integrals along any k-chain σ. Consequently, these two k-forms coincide. The homotopy formula is proved. □

Remark. Our previous "co-cone" proof of the Poincaré lemma (p. 219) was based on a similar reasoning, and may be deduced from the present one, applied to the special vector field $v(x) = x$ in $\mathbb{R}^n$.

Problem. Prove that

$$L_v(\alpha \wedge \beta) = (L_v\alpha) \wedge \beta + \alpha \wedge (L_v\beta). \tag{5}$$

Solution. Differentiating the exterior product $\alpha \wedge \beta$ in the parameter t, we write

$$(\alpha + \alpha') \wedge (\beta + \beta') = (\alpha \wedge \beta) + \alpha' \wedge \beta + \alpha \wedge \beta' + O(|\alpha'|^2 + |\beta'|^2),$$

which provides formula (5) for small increments α' and β'.

Problem. Prove that L_v and d commute: $L_v(d\omega) = d(L_v\omega)$.

Solution. The fact that the exterior derivative operation d is independent of the choice of coordinates means that it commutes with diffeomorphisms (since the diffeomorphisms can be treated as changes of coordinates). The proof follows because the equality $Ld = dL$ is an infinitesimal version of the mentioned above property, when the diffeomorphism is the phase flow of the field for small time.

A more formal algebraist mind could prefer the following calculations:

$$L_v(d\omega) = i_v d(d\omega) + d(i_v(d\omega)) = d(i_v(d\omega)),$$

$$d(L_v\omega) = d(i_v(d\omega) + dd(i_v\omega)) = d(i_v(d\omega)),$$

which hold since $d^2 = 0$ both for ω and for $i_v\omega$.

7.3.2 Hamilton Vector Fields in Symplectic Spaces

Problem. Find the vector fields that preserve the (symplectic) 2-form

$$\omega = dp_1 \wedge dq_1 + \ldots + dp_n \wedge dq_n$$

in the (symplectic) space $\mathbb{R}^{2n}$, equipped with Darboux coordinates (p_k, q_k), $k = 1, \ldots, n$. (see § 3.15, p.107. Symplectic geometry is discussed in Ch. 16.)

Solution. The condition of the preservation of the 2-form ω, $L_v\omega = 0$, can be written, according to the homotopy formula, as the equality

$$i_v(d\omega) + d(i_v\omega) = 0\,.$$

The first term vanishes, since $d\omega = 0$. Now, by the Poincaré lemma, the condition $d(i_v\omega) = 0$ implies that the differential 1-form $i_v\omega$ is the differential of some function. We denote $-H$ this function, following the traditional notations in physics where the function H is called the *Hamilton function* of the field v and the field v is called the *Hamilton vector field* of H.

Writing explicitly $i_v\omega$ for the symplectic 2-form $\omega = \sum_{k=1}^n dp_k \wedge dq_k$ and the vector field $v = \sum_{k=1}^n (A_k\ \partial/\partial p_k + B_k\ \partial/\partial q_k)$ we obtain the 1-form $i_v\omega = \sum_{k=1}^n (A_k dq_k - B_k dp_k)$ in the Darboux coordinates.

The equality $i_v\omega = -dH$, for the Hamilton vector field v with Hamilton function H, provides the components $A_k = -\partial H/\partial q_k$, $B_k = \partial H/\partial p_k$. Hence

$$v = \sum_{k=1}^{n} \left(\frac{\partial H}{\partial p_k}\frac{\partial}{\partial q_k} - \frac{\partial H}{\partial q_k}\frac{\partial}{\partial p_k} \right)$$

and the resulting Hamilton differential equation $\dot{x} = v(x)$ takes the form

$$\dot{p}_k = -\partial H/\partial q_k, \qquad \dot{q}_k = \partial H/\partial p_k,$$

which explains the choice of the sign, $i_v\omega = -dH$.

Example. Consider the torus surface $\mathbb{T}^2$ equipped with the Darboux angular coordinates $p \pmod{2\pi}$ and $q \pmod{2\pi}$ and with the symplectic 2-form $\omega = dp \wedge dq$. The "rigid motions" $g^t(p,q) = (p + \alpha t, q + \beta t)$ on $\mathbb{T}^2$ (with angular velocities α and β) are symplectomorphisms (i.e. they preserve the symplectic 2-form) and they form a one-parameter group, which is the phase flow of its vector field of velocities

$$v = \alpha\ \partial/\partial p + \beta\ \partial/\partial q.$$

Problem. Is v a Hamilton vector field on the torus?

Answer. No, it is not. Indeed, $i_v\omega = \alpha dq - \beta dp$, and this (closed) differential 1-form on $\mathbb{T}^2$ is the differential of a function $-H$ on $\mathbb{T}^2$ only in the case $\alpha = \beta = 0$.

So, the velocity fields of the families of symplectomorphisms are in general non-Hamiltonian vector fields. The vanishing of the differential $d(i_v\omega)$ of the form $i_v\omega$ does not imply that this closed 1-form is a differential. As we explained in p. 219, this reflects the topology of the ambient manifold.

In this sense, the vector fields defined by the velocities of change of a symplectomorphism (with respect to time) are, in general, only *locally Hamiltonian* vector fields. Globally, their Hamilton functions are multivalued, as for the above "function" $\alpha q - \beta p$ on the torus $\mathbb{T}^2$.

7.4 Poisson (Lie) Bracket of Vector Fields

The fisherman derivative $L_v w$ of a vector field w on a smooth manifold M along another vector field v on M, provides an interesting vector field that we shall describe.

Theorem 1. *The Lie derivative of a smooth vector field* $w = \sum w_k \, \partial/\partial x_k$ *along a smooth vector field* $v = \sum v_k \; \partial/\partial x_k$ *has* k*th component* $L_w v_k - L_v w_k$ *:*

$$L_v w = \sum (L_w v_k - L_v w_k) \, \frac{\partial}{\partial x_k} \, . \tag{6}$$

Proof. We shall use the fact that the value of a differential form on a vector is intrinsically defined, independently of any choice of the coordinate system, and is therefore preserved by any flow consisting of diffeomorphisms.

Let α be a smooth differential 1-form on M and w be a smooth vector field on M. Then the value

$$i_w \alpha = \alpha(w) = f$$

is a smooth function on M, and this bilinear relation between α, w and f is preserved by the diffeomorphisms g^t which form the flow of the field v. Hence,

$$(g^{t^*} \alpha)(g_*^{-t} w) = g^{t^*} f \tag{7}$$

(we take $-t$ to move the field w back, like the form α and the function f).

We derivate in t both sides of this relation (at $t = 0$), and we use the definition of the fisherman derivative to get:

$$\frac{d}{dt}|_{t=0}(g^{t^*} \alpha) = L_v \alpha, \quad \frac{d}{dt}|_{t=0}(g_*^{-t} w) = -L_v w, \quad \frac{d}{dt}|_{t=0}(g^{t^*} f) = L_v f.$$

Therefore, the derivation of identity (7) provides the identity

$$\alpha(-L_v w) + (L_v \alpha)(w) = L_v(\alpha(w)). \tag{8}$$

This identity provides the explicit formulas for the components of the vector field $L_v w$. Indeed, in formula (8) we can take as α the differential dx_k of the coordinate function x_k, whose value on any vector field is the k-th component of that field in our coordinated system. So, the k-th component of the derivated vector field $L_v w$ has the form

$$dx_k(-L_v w) = L_v((dx_k)(w)) - (L_v(dx_k))(w).$$

The homotopy formula provides the following expression for the last term

$$L_v(dx_k) = i_v d(dx_k) + d(i_v(dx_k)) = dv_k,$$

where $v_k = i_v(dx_k)$ is the k-th component of the vector field v.

Writing w_k for the components of the vector field $w = \sum w_k\ \partial/\partial x_k$, the k-th component of the derivated vector field $L_v w = a_k\ \partial/\partial x_k$ is, then, given by the expression:

$$a_k = (dv_k)(w) - L_v w_k = L_w v_k - L_v w_k\,. \qquad \square$$

In the following theorems and problems the linear differential operators will be denoted with capital letters, while the vector fields with small letters.

Consider the linear homogeneous operators

$$V = \sum v_k\ \partial/\partial x_k \quad \text{and} \quad W = \sum w_k\ \partial/\partial x_k$$

defined by the respective vector fields v and w.

Theorem 2. *The commutator $VW - WV$ of these two first order differential operators is itself a first order linear differential operator.*

(Therefore, the operator $A = VW - WV$ is the derivation (of functions) along a vector field a, in spite of the fact that it seems to include the second derivatives!)

Proof. Applying the product of the operators V and W to a smooth function f, we obtain

$$VWf = \sum_\ell v_\ell \frac{\partial}{\partial x_\ell}\left(\sum_k w_k \frac{\partial f}{\partial x_k}\right) = \sum_{k,\ell}\left(v_\ell \frac{\partial w_k}{\partial x_\ell}\frac{\partial f}{\partial x_k} + v_\ell w_k \frac{\partial^2 f}{\partial x_\ell \partial x_k}\right).$$

The similar expression for WVf contains the term $w_k v_\ell \frac{\partial^2 f}{\partial x_k \partial x_\ell}$. However, since the second partial derivatives are symmetric, the difference $VW - WV$ contains no second derivative at all:

$$(VW - WV)f = \sum_k (L_v w_k - L_w v_k)\,\frac{\partial f}{\partial x_k}\,. \qquad \square$$

Formula (6) of p.241 leads to the following conclusion:

Corollary. *The vector field that defines the differential operator $VW - WV$ is the field $-L_v w$:* the fisherman derivative of w along v (with a minus sign).

Poisson bracket. Given two operators V and W (defined by the respective vector fields v and w), the vector field a that represents the commutator $A = VW - WV$ is called the *Poisson bracket* of the vector fields v and w. It is denoted by $a = \{v, w\}$, that is

$$\{v, w\} = -L_v w.$$

Exercise. Prove the following (evident) properties of the Poisson bracket
(i) $\{v, fw\} = f\{v, w\} + (L_v f)\, w$,
(ii) $\{u + v, w\} = \{u, w\} + \{v, w\}$ (and of course $\{w, v\} = -\{v, w\}$).

Problem. Let v be the vector field of the velocities of the points of a rigid body which rotates with angular velocity α, and w be the velocity field defined by another body rotating with angular velocity β (we suppose that both rotation axes contain the origin of Euclidean space $\mathbb{R}^3$). Find the Poisson bracket $\{v, w\}$ of the fields of velocities.

Answer. It is the velocity field defined by the rotation of a rigid body with angular velocity $[\alpha, \beta]$ (the vector product in oriented Euclidean space $\mathbb{R}^3$).

Solution. The velocity fields are $v(x) = [\alpha, x]$ and $w(x) = [\beta, x]$. Thus, we obtain the linear expression $v(x) = Ax$, where (A) is the antisymmetric matrix $(A) = \begin{pmatrix} 0 & -p & r \\ p & 0 & -q \\ -r & q & 0 \end{pmatrix}$, for $\alpha = (p, q, r)$. Similarly $w(x) = Bx$, where (B) is the antisymmetric matrix constructed from the angular velocity β.

The calculation of the Poisson bracket is now reduced to the calculation of the difference $AB - BA$, which is also antisymmetric and is hence equal to the velocity field of a rotating body with the corresponding angular velocity.

Multiplying the matrices A and B, we get the angular velocity $[\alpha, \beta]$.

Problem. Let v be a linear vector field $v(x) = Ax$ in $\mathbb{R}^n$ defined by a linear operator $A : \mathbb{R}^n \to \mathbb{R}^n$, that is, $v(x) = \sum a_{ij} x_j \partial/\partial x_i$. Similarly $w(x) = Bx$. Calculate the Poisson bracket $\{v(x), w(x)\}$ of these two vector fields.

Answer. It is the linear vector field $u(x) = [B, A]x$, where $[B, A] = BA - AB$ is the commutator of the matrices B and A, that is, $\{Ax, Bx\} = [B, A]x$.

Problem. Prove that the commutators $\{A, B\} = AB - BA$ verify the following Jacobi identity:

$$\{\{A, B\}, C\} + \{\{B, C\}, A\} + \{\{C, A\}, B\} = 0.$$

Solution. The three double commutators provide 12 triple products of the multiplied operators A, B and C in different orders and with different signs. Only 6 different permutations of 3 objects A, B, C are possible, and each one is presented twice with opposite signs. Hence, the sum vanishes.

Problem (Jacobi Identity). Prove that the identity

$$\{\{u,v\},w\}+\{\{v,w\},u\}+\{\{w,u\},v\}=0.$$

holds for any smooth vector fields u, v, w.

Solution. The Jacobi identity for the corresponding first order linear differential operators U, V, W, it is just the identity of the preceding problem, whose solution, hence, provides that property for vector fields.

7.4.1 Poisson Bracket in Symplectic Spaces

Problem. Prove that *the Poisson bracket of two Hamilton vector fields is also a Hamilton vector field*.

Solution. Consider two Hamilton vector fields v, w with respective Hamilton functions H, F. We shall derivate the defining relation $i_w\omega = -dF$ along the Hamilton vector field v of the Hamilton function H (that is, $i_v\omega = -dH$).

Since the symplectic form ω is invariant under the flow of the Hamilton field v, the derivative of the left hand side of the equality $i_w\omega = -dF$ is

$$L_v(i_w\omega) = i_{L_v w}(\omega),$$

while the derivative of the right hand side is (see p. 239)

$$-L_v(dF) = -d(L_vF).$$

This proves that the Poisson bracket $\{w, v\} = L_vw$ is the Hamilton vector field of the Hamilton function $L_vF = i_v(dF)$.

Poisson Bracket of Functions. The Hamilton function $L_vF = i_v(dF)$ is called *Poisson bracket of the functions* H *and* F. It is denoted by $\{H, F\}$.

Exercise. For two Hamilton functions H, F with respective Hamilton fields v, w, write the Hamilton function $\{H, F\}$ in Darboux coordinates (p_k, q_k).

ANSWER. In Darboux coordinates $v = \sum_{k=1}^{n} \left(\frac{\partial H}{\partial p_k} \frac{\partial}{\partial q_k} - \frac{\partial H}{\partial q_k} \frac{\partial}{\partial p_k} \right)$, see p. 240. Since the Hamilton function $\{H, F\}$ of the vector field $L_v w$ equals $L_v F$, we have

$$\{H, F\} = \sum_{k=1}^{n} \left(\frac{\partial H}{\partial p_k} \frac{\partial F}{\partial q_k} - \frac{\partial H}{\partial q_k} \frac{\partial F}{\partial p_k} \right). \tag{9}$$

Theorem 3. *The Poisson brackets of smooth functions on a symplectic manifold verify the Jacobi identity,*

$$\{\{F, G\}, H\} + \{\{G, H\}, F\} + \{\{H, F\}, G\} = 0.$$

Proof. Although this identity was already proved for the Hamilton vector fields, it remains the possibility to have at the right hand side any constant (whose Hamilton vector field would vanish). The fact that this constant is zero follows from the representation of the differential operators

$$\{H, \{F, G\}\} = L_v\{F, G\} = L_v(L_w G), \qquad \{F\{G, H\}\} = -L_w(L_v G)$$

and $$\{G, \{H, F\}\} = L_a G,$$

where a is the Hamilton field of the Hamilton function $\{H, F\}$.

As we have proved above, the field a is (minus) the Poisson bracket of the fields v and w. So, $L_a = L_w L_v - L_v L_w$, and therefore

$$\begin{aligned}\{H, \{F, G\}\} + \{F, \{G, H\}\} + \{G, \{H, F\}\} &= (L_v L_w - L_w L_v + L_w L_v - L_v L_w)G \\ &= 0,\end{aligned}$$

which proves the Jacobi identity. □

PROBLEM. Calculate the Poisson bracket of any two quadratic forms in the symplectic space $\mathbb{R}^{2n}$.

ANSWER. Formula (9) above shows that the Poisson bracket of two quadratic forms is still a quadratic form (it also shows that the Poisson bracket of two forms of degrees k and l is a form of degree $k + l - 2$).

Example. For $n = 1$ the space of quadratic forms is 3-dimensional: All such forms have the expressions

$$\alpha p^2 + \beta pq + \gamma q^2,$$

with constant coefficients α, β, γ.

The Poisson bracket operation acts in this 3-dimensional space as a bilinear anti-symmetric operation $\mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$.

Surprisingly, the linear algebra of this symplectic Poisson bracket operation describes the geometry of the Lobachevsky plane (together with its continuation to the relativistic de Sitter's world geometry in the Möbius band, which is the complement to the Lobachevsky disc in $\mathbb{RP}^2$).

For instance, in these geometries the Jacobi identity has the following geometric interpretation: It is just the theorem on the concurrence of the altitudes, which states that the three altitude lines of a triangle in Lobachevsky plane always have a common point (which sometimes belongs, however, to the de Sitter complementary relativistic world, rather than to Lobachevsky plane). In next chapter we study Lobachevsky plane geometry.

The details of these relations between the Poisson bracket of quadratic forms on the symplectic plane and the geometries of Lobachevsky and of de Sitter are published in [35].

7.5 Jacobi Identity - triangle altitudes theorem

Problem. Write $[v, w]$ for the vector product of two vectors v and w in oriented Euclidean 3-space. Prove the Jacobi identity

$$[[u, v], w] + [[v, w], u] + [[w, u,], v] = 0$$

for any three vectors u, v, w.

Solution. This Jacobi identity follows from the interpretation of the vector product as the Poisson bracket of the velocity fields of the rigid rotations (with angular velocities u, v and w), see the problem on page 243.

This identity has an elementary meaning in plane Euclidean geometry :

Theorem 4 (proved below)**.** *The three altitudes of a triangle in a Euclidean plane have a common intersection point.*

The proof use the following five elementary lemmas on the vector product.

Consider a triangular pyramidal cone in oriented Euclidean space $\mathbb{R}^3$. Write U, V, W for the three planes which form the faces of the pyramidal cone, and write u, v, w for the three vectors orthogonal to the faces at the vertex 0 of the pyramid – Fig. 7.8.

$U \,:\, BOC$ $\qquad P_w \,:\, COc$

$V \,:\, COA$ $\qquad R \parallel Q$

$W \,:\, AOB$ $\qquad q \perp R$

$R \,:\, ABC$ $\qquad p_w \perp P_w$

$p_w = [[u, v], w]$ $\qquad q \perp Q$

$w \perp W$

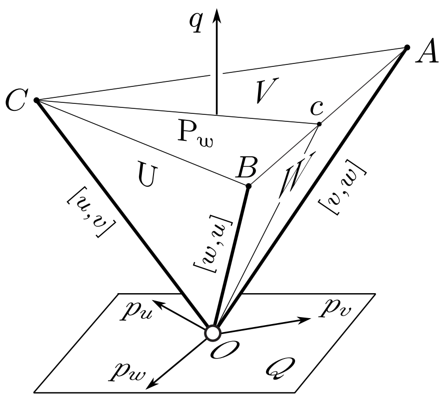


Figure 7.8: The altitudes theorem of Euclidean geometry from the Jacobi identity $p_u + p_v + p_w = 0$.

Lemma 1. *The vector* $[u, v]$ *belongs to the intersection* $U \cap V$.

Proof. The vector product is orthogonal to both multiplied vectors. $\square$

Lemma 2. *The vector* $p_w = [[u, v], w]$ *lies in the plane* W.

Proof. It is orthogonal to w, which is orthogonal to the plane W. $\square$

Lemma 3. *The vector* p_w *is orthogonal to the edge* $\mathbb{R}[u, v]$ *of the pyramid* (opposite to the face W).

Proof. This follows again from orthogonality of the vector-product to each multiplied vector. $\square$

Lemma 4. *The plane* P_w*, orthogonal to the vector* p_w *at* O*, is the altitude plane of the pyramid* (it is orthogonal to the plane W and it contains the opposite edge $\mathbb{R}[u, v]$).

Proof. The edge is contained in P_w, by Lemma 3, and the orthogonality to the plane W follows from the fact that $w \in P_w$. $\square$

Lemma 5. *The three altitude planes of the pyramid,* P_u*,* P_v*,* P_w*, have a common line.*

Proof. Their three orthogonal vectors p_u, p_v, p_w have the sum equal to zero, by the Jacobi identity, and hence these three vectors belong to a common plane Q: $p_u \in Q$, $p_v \in Q$, $p_w \in Q$.

The vector q, orthogonal to the plane Q at O, is orthogonal to each of the three vectors p_u, p_v, p_w, and belongs therefore to the intersection of the three planes: $q \in P_u \cap P_v \cap P_w$.

Thus the three altitude planes have a common straight line $\mathbb{R}q$. □

Proof of Theorem 4. Consider a plane R orthogonal to the line $\mathbb{R}q$ at some point different from O. It intersects the faces of the pyramid along a triangle ABC (see Fig. 7.8). We shall prove that the three altitude planes (P_u, P_v, P_w) intersect the plane R along the altitudes $(Cc\dots)$ of the triangle ABC.

The side AB of the triangle is parallel to the vector p_w, since the intersections of the plane W with the parallel planes Q and R are parallel lines. The line Cc (of intersection of the plane P_w with the plane R) is orthogonal to the vector p_w (which is orthogonal to the plane P_w, by the definition of this plane). Thus, the line Cc is orthogonal to the side AB of the triangle ABC, that is, Cc is an altitude of the triangle ABC. We proceed similarly for the lines Aa and Bb.

We have proved that the three altitudes (Cc, Bb, Aa) of the triangle ABC intersect each other (at their common point $tq \in R$).

The triangle ABC may be any triangle (for a suitable choice of the three planes U, V, W). □

Remark. We have proved that the theorem of Euclidean geometry on the concurrence of the altitudes of a triangle is a geometric corollary of the algebraic Jacobi identity. Our reasoning could be inverted to deduce the Jacobi identity (for the vector product) from the geometric theorem on the altitudes. However, in this way we obtain only the weaker statement that the three vectors p_u, p_v and p_w are linearly dependent (instead of the Jacobi identity claim, $p_u + p_v + p_w = 0$).

It would be interesting to understand the geometric meaning of this additional statement to the altitudes theorem, but we have not found such geometrical interpretation of the Jacobi identity.

7.6 Lie groups and Lie algebras

In Chapter 3, we have seen that the manifold of rotations $\mathrm{SO}(3)$ is diffeomorphic to $\mathbb{R}\mathrm{P}^3$, the manifold of unit quaternions $\mathbb{S}^3 = \mathrm{Spin}(3)$ coincides with the group $\mathrm{SU}(2)$ and the unitary group $\mathrm{U}(n)$ coincides with the three pairwise intersections of the three subgroups (which are also submanifolds) $\mathrm{GL}(n,\mathbb{C})$, $\mathrm{O}(2n)$ and $\mathrm{Sp}(n,\mathbb{R})$ of $\mathrm{GL}(2n,\mathbb{R})$.

We have also seen the isomorphism $\pi_1(\mathrm{SO}(n)) \approx \mathbb{Z}_2$ for $n \geqslant 3$, the homotopy group isomorphisms $\pi_k(\mathrm{GL}^+(n,\mathbb{R})) \approx \pi_k(\mathrm{SL}(n,\mathbb{R})) \approx \pi_k(\mathrm{SO}(n,\mathbb{R}))$ and the Bott periodicity relations of the "stable homotopy groups" $\pi_k(\mathrm{SL}(\infty)) \approx \pi_k(\mathrm{SO}(\infty)) \approx \pi_{k+8}(\mathrm{SO}(\infty))$.

All these cases are about groups that are also smooth manifolds. We shall see that the interaction of the manifold and group structures provides lots of interesting results.

Lie groups. A *Lie group* G is a smooth manifold which is also a group and such that the group operation $G \times G \to G$ is a smooth map, and that the inverse operation $G \to G$ (sending g to g^{-1}) is a diffeomorphism.

Examples of Lie Groups.

1. The spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ with addition operation are abelian Lie groups.

2. The unit circle $\mathbb{S}^1 = \{z \in \mathbb{C} :| z |= 1\}$ with multiplication operation is a an abelian Lie group.

3. The orthogonal groups $\mathrm{O}(n)$, $\mathrm{SO}(n)$, the unitary groups $\mathrm{U}(n)$, $\mathrm{SU}(n)$, the symplectic groups $\mathrm{Sp}(n)$, the groups $\mathrm{Spin}(n)$, the general linear groups $\mathrm{GL}(n,\mathbb{R})$, $\mathrm{GL}(n,\mathbb{C})$ and the determinant 1 "special" linear groups $\mathrm{SL}(n,\mathbb{R})$, $\mathrm{SL}(n,\mathbb{C})$ are Lie groups (mostly non commutative, like $\mathrm{SO}(3)$, $\mathrm{U}(2)$, $\mathbb{S}^3$).

4. The projective transformations of the real projective line $\mathbb{R}\mathrm{P}^1 \approx \mathbb{S}^1$, whose expression is $x \mapsto \frac{ax+b}{cx+d}$, form a 3-dimensional Lie group which is denoted

$$\mathrm{PL}(2,\mathbb{R}) = \frac{\mathrm{SL}(2,\mathbb{R})}{\{\pm 1\}}.$$

Similarly, the projective transformations of the complex projective line form the Lie group

$$\mathrm{PL}(2,\mathbb{C}) = \frac{\mathrm{SL}(2,\mathbb{C})}{\{\pm 1\}},$$

which is a smooth complex manifold of complex dimension 3.

5. The affine transformations of the line $\mathbb{R} = \{x\}$, $x \mapsto ax + b$ $(a \neq 0)$, form a Lie group denoted $\mathrm{Aff}(1,\mathbb{R})$. The manifold $\{(a,b) : a \neq 0\}$ consists of two half-planes, $a > 0$ ("upper half-plane") and $a < 0$.

The upper half-plane (which contains the unity $(a = 1, b = 0)$) forms itself a Lie group. It is denoted $\mathrm{Aff}^+(1,\mathbb{R})$ and is a subgroup of the affine group $\mathrm{Aff}(1,\mathbb{R}) = \{a, b : a \neq 0\}$. The group $\mathrm{Aff}^+(1,\mathbb{R})$ provides a natural geometric description of the Lobachevsky plane (see Chapter 8).

We shall see below that the tangent space of a Lie group G at the unit element e, $T_eG \approx \mathbb{R}^{\dim G}$, is naturally equipped with an operation ("Lie bracket") which is an infinitesimal version of the group operation on the manifold G.

7.6.1 Left- and Right-invariant Vector Fields

Definition. The *left translation* on a Lie group G by its element g is the diffeomorphism $L_g : G \to G$ provided by the left multiplication of all the elements of G by the fixed element g. That is, $L_g h = gh$ for any $h \in G$.

Left Invariant Fields. A vector field v on a Lie group G is called *left invariant* if all left translations send it to itself:

$$(L_g)_* v = v, \qquad (L_g)_*(v(x)) = v(L_g x),$$

where $v(x) \in T_x G$. A left invariant vector field is completely determined by its value at one point, say, at e:

$$v(g) = L_{g*}(v(e)),$$

and $v(e) \in T_e G$ may be chosen arbitrarily.

Similarly, the right translation $R_g : G \to G$ is the diffeomorphism defined as the right multiplication by $g \in G$. That is, $R_g h = hg$ for any $h \in G$.

Right Invariant Fields. The *right invariant* vector fields are defined similarly

$$R_{g*} v = v, \qquad (R_g)_*(v(x)) = v(R_g x)\,.$$

A right invariant vector field is completely determined by its value at e:

$$v(g) = R_{g*}(v(e)).$$

The General Linear Group. Consider the Lie group $\mathrm{GL}(n, \mathbb{R})$ of the non-degenerate real matrices of order n. Its tangent space (of dimension n^2) consists of all the $n \times n$ real matrices (degenerate or not).

Problem. Let v and w be the two vector fields on $\mathrm{GL}(n, \mathbb{R})$ whose values at the point 1 (the identity matrix) are two given matrices A and B. Prove that if v and w are right invariant fields then their Poisson bracket is also a right invariant field and compute its value C at the point 1.

Answer. The right invariance follows from the intrinsic nature of the Poisson bracket, which is invariant under diffeomorphisms:

$$\{g_* v, g_* w\} = g_* \{v, w\}\,,$$

where the natural action g_* of a diffeomorphism g on a vector field v is defined by the formula $(g_* v)(g(x)) := (g_{*x}) v(x)$ – Fig. 7.9.

The value of the Poisson bracket at the point 1 is “minus the commutator” $C = BA - AB$ of the matrices A and B.

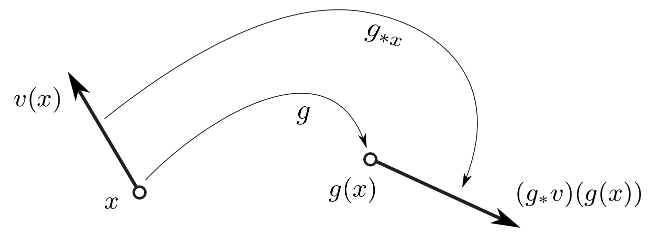


Figure 7.9: The action g_* of a diffeomorphism g on a vector field v.

PROBLEM. Make the corresponding calculations for the left invariant fields.

ANSWER. For left invariant fields it is the commutator $C = AB - BA$.

SOLUTION. Both formulas follow from the explicit calculation of the fields: Our group G is "locally" covered by a coordinate system $\mathbb{R}^{n^2}$ because the manifold of non-degenerate matrices is an open domain in the space of all $n \times n$ real matrices. Using this coordinate system, the tangent vectors of the group are also denoted by matrices (we identify $T_x\mathbb{R}^{n^2}$ with the space $\mathbb{R}^{n^2}$).

In the left invariance case, a vector A at 1 (which is the velocity vector of the motion $1+At+\ldots$) is sent by the left translation L_g to the velocity vector of the motion $L_g(1+At+\ldots) = g+gAt+\ldots$, which is $v(g) = gA$. Similarly, in the right invariance case, the right-translated vector is the velocity vector of the motion $R_g(1+At+\ldots) = g+Agt+\ldots$, which is $v(g) = Ag$.

At each point g of our group, the Poisson bracket of the right invariant vector fields $v(g) = Ag$ and $w(g) = Bg$ is given by the standard formula

$$\{v, w\}(g) = \frac{\partial w}{\partial g}v - \frac{\partial v}{\partial g}w = BAg - ABg = (BA - AB)g.$$

Similarly, for the left invariant fields $v(g) = gA$ and $w(g) = gB$ we have

$$\{v, w\}(g) = v\frac{\partial w}{\partial g} - w\frac{\partial v}{\partial g} = gAB - gBA = g(AB - BA).$$

PROBLEM. Let A and B be the initial velocities of two motions of the identity matrix 1: $a(t) = 1 + At + o(t)$ and $b(s) = 1 + Bs + o(s)$. Calculate the quadratic term of the Taylor series, at $s = t = 0$, of the matrix

$$C(s,t) = a(t)b(s)(a(t))^{-1}(b(s))^{-1}.$$

Solution. Since $C(0,t) = C(s,0) = 1$, the only second order term is bilinear in s and t. Taking this into account, the second order terms s^2 and t^2 will be neglected in our calculations. We first compute the two products

$$(1 + At + o(t))(1 + Bs + o(s)) = 1 + At + Bs + ABts + o(ts)\,,$$

$$(1 - At + o(t))(1 - Bs + o(s)) = 1 - At - Bs + ABts + o(ts)\,.$$

Then, we multiply the right hand sides of these equalities to get

$$C(s,t) = 1 + (AB - BA)st + o(st).$$

We have solved the problem and additionally proved the important

Theorem 5. *The principal (bilinear) term of the deviation of the product $C(s,t)$ from the identity matrix 1 is equal to $st\,(AB - BA)$.*

The geometric meaning of this theorem is that the commutator tangent vector $\{A,B\} := AB - BA$ is the main part of the vector that joins the end points of the curvilinear pseudo-parallelogram constructed from the phase curves of the right multiplications by $a(t)$, then by $b(s)$, and then returning back dividing by $a(t)$ and by $b(s)$ – Fig. 7.10.

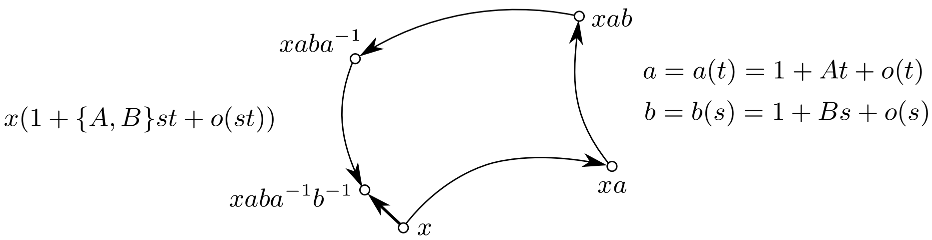


Figure 7.10: The infinitesimal displacement $\{A,B\}st$, of the unity matrix, produced by the infinitesimal parallelogram with sides At and Bs.

In this sense, the commutator $\{A,B\}$ "measures" the non-commutativity of the motions a and b, in terms of their velocities A and B.

Lie bracket. This bilinear operation $A, B \mapsto \{A,B\}$ in the tangent space at the point e is defined for any Lie group in the same ways as in the examples above: Either as the Poisson bracket operation between the right invariant vector fields, or as the coefficient of the bilinear part of the deformation $C(s,t)$ of the unity element. Both definitions determine the same bilinear map $\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$, of the tangent space $\mathfrak{g} = T_e G$ to itself, which is called the *Lie bracket*: $(A \in \mathfrak{g}, B \in \mathfrak{g}) \mapsto ([A, B] \in \mathfrak{g})$.

This operation is always anti-symmetric, $[B, A] = -[A, B]$ and verifies the Jacobi identity

$$[[A, B], C] + [[B, C], A] + [[C, A], B] = 0.$$

Lie algebra. A *Lie algebra* is a vector space equipped with a bilinear anti-symmetric binary operation verifying the Jacobi identity (*Lie bracket*).

7.6.2 The Lie Algebra of a Lie Group

Since the tangent space of a Lie group G at the unity, $T_e G$, has a natural Lie algebra structure, it is called *the Lie algebra space of that group.* It is usually noted with the corresponding lower case Gothic letter $\mathfrak{g}$.

Much of the structure of a Lie Group is encoded in its Lie algebra, which is a more simple algebraic object. In the finite dimensional case, a connected simply connected Lie Group can be reconstructed from its Lie algebra. Moreover, the Lie subalgebras of the Lie algebra $T_e G$ are in one-to-one correspondence with the Lie subgroups of G.

We state without proof the following two theorems.

Theorem (E. Cartan)**.** *Every finite dimensional Lie algebra is the Lie algebra space of a suitable Lie group.* (In the infinite-dimensional case this is not so.)

Ado's Theorem (see [73]). *Every finite dimensional Lie algebra is isomorphic to a Lie subalgebra of* $\mathfrak{gl}(n)$.

Exercise. Classify all one-dimensional Lie algebras.

Answer. There is only one 1-dimensional Lie algebra: the real vector line $\mathbb{R}$ with (obviously) null Lie bracket, $[\ ,\ ] = 0$.

Therefore the additive group $\mathbb{R}$ and the circle group SO(2), both one-dimensional Lie groups, have isomorphic Lie algebra spaces.

Problem. Study the Lie algebras of the Lie groups SO(3) and $\mathbb{S}^3$ (the group of quaternions of norm 1). Are these two Lie algebras isomorphic?

Answer. Both Lie bracket operations are isomorphic to the vector product operation in oriented Euclidean 3-space $\mathbb{R}^3$.

7.6.3 Infinitesimal Generators of Group Actions

When a Lie group acts (by diffeomorphisms) on a manifold M, its Lie algebra is represented by some vector fields on M (the infinitesimal generators of the group of diffeomorphisms). Then the Lie bracket operation on the Lie algebra becomes the Poisson bracket operation on these vector fields.

PROBLEM. Calculate the Lie algebra structure of the Lie group formed by the projective transformations of the real projective line.

SOLUTION. The projective map $x \mapsto \frac{ax+b}{cx+d}$, corresponds to the matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. For $a = 1 = d$, $b = c = 0$ it is the identity map $x \mapsto x$; deforming it infinitesimally by a vector $\left(\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right)$ of the Lie algebra $\mathfrak{g}$ we get $\left(\begin{smallmatrix} 1+\alpha & \beta \\ \gamma & 1+\delta \end{smallmatrix}\right)$.

Since the vector $\left(\begin{smallmatrix} \alpha & 0 \\ 0 & 0 \end{smallmatrix}\right) \in \mathfrak{g}$ produces the map $x \mapsto (1+\alpha)x = x + \alpha\, x$, it is represented by the vector field $\alpha\, x\, \frac{\partial}{\partial x}$.

The vector $\left(\begin{smallmatrix} 0 & 0 \\ \gamma & 0 \end{smallmatrix}\right) \in \mathfrak{g}$ produces the map $x \mapsto \frac{x}{\gamma x+1} = (x - \gamma x^2 + \ldots)$ and therefore it is represented by the vector field $-\gamma\, x^2\, \frac{\partial}{\partial x}$.

The reader should check that the vectors $\left(\begin{smallmatrix} 0 & \beta \\ 0 & 0 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & 0 \\ 0 & \delta \end{smallmatrix}\right)$ are represented by the respective vector fields $\beta\, \frac{\partial}{\partial x}$ and $-\delta\, x\, \frac{\partial}{\partial x}$.

Therefore, the generators of the vector fields representing the Lie algebra are the three vector fields (found on p. 232): $\frac{\partial}{\partial x}\,,\ x\frac{\partial}{\partial x}\,,\ x^2\frac{\partial}{\partial x}\,.$

These three vector fields can also be found via the Taylor series of a smooth function

$$f(x) \mapsto f\left(\frac{(1+\alpha)x+\beta}{\gamma x + (1+\delta)}\right) = f(x) + \beta f'(x) + (\alpha - \delta) x f'(x) - \gamma x^2 f'(x) + \ldots\,.$$

Calculate the Poisson brackets of the three vector fields $x^n\ \partial/\partial x$ ($n = 0, 1, 2$). The smooth vector fields on the projective line which are defined in the affine coordinate x by the 3-parameter formula

$$v = (c_0 + c_1 x + c_2 x^2)\frac{\partial}{\partial x},$$

are called *Riccati vector fields*. They are the velocity fields of all one-parameter groups of projective transformations of the real projective line.

PROBLEM. Calculate the Lie algebra structure of the group of affine transformations of the affine real line $\mathbb{R}$.

ANSWER. In our case:

$$A = \frac{\partial}{\partial x}, \qquad B = x\frac{\partial}{\partial x}, \qquad [A, B] = A.$$

Problem. Find the Poisson bracket of a right invariant vector field v on a Lie group G, with a left invariant vector field w on G.

Solution. The left translations on G commute with the right translations: $L_g R_h = R_h L_g$, since

$$L_g(R_h k) = L_g(kh) = gkh \quad \text{and} \quad R_h(L_g k) = R_h(gk) = gkh\,.$$

In consequence, the left invariant fields commute with the right invariant fields: $\{v, w\} = 0$.

Problem. Find the Lie algebra of the special orthogonal group $\mathrm{SO}(n)$.

Answer. It is formed by the anti-symmetric real matrices, i.e. $A + A^T = 0$, with Lie bracket $[A, B] = AB - BA$. *Hint.* For $A \in \mathfrak{g}$, take a curve $g(t)$ in $\mathrm{SO}(n)$ with $\dot{g}(0) = A$ at $g(0) = e$ and derive the equality $g(t)\, g(t)^T = e$ at $t = 0$.

Problem. Find the Lie algebra of the group $\mathrm{U}(n)$ of unitary matrices.

Answer. It is formed by the anti-Hermitian matrices of order n, $A^* = -A$ (the operation * sends the complex matrix $(a_{j,k})$ to the complex matrix $(\overline{a_{k,j}})$, where the bar denotes the complex conjugation operation: $\overline{x + iy} = x - iy$). This Lie algebra is a *real* vector space of dimension n^2 (in spite of the fact that the matrices forming it have complex elements).

The Lie bracket $[A, B] = AB - BA$ is still anti-Hermitian, since

$$(AB - BA)^* = B^*A^* - A^*B^* = BA - AB\,.$$

7.7 Lie Bracket and Adjoint Representation

In § 2.4, we have defined an "action of a group G on a given set V" as a homomorphism of the group G into the group of transformations of V.

A *representation* of G is a group action in which V is a vector space and the transformations τ_g are invertible linear maps, that is, $\tau : G \to \mathrm{GL}(V)$.

In the case of a Lie group G, a representation $\tau : G \to \mathrm{GL}(V)$ is a homomorphism between Lie groups which is also a smooth map between the underlying manifolds. Its differential at the unity is then a *morphism of Lie algebras* (it preserves the Lie bracket), $d_e\tau : T_eG \to T_0\mathrm{GL}(V) \simeq \mathrm{End}\,(V)$.

Every Lie group G has an intrinsic representation on its own Lie algebra T_eG. The corresponding morphism of Lie algebras $T_eG \to \mathrm{End}\,(T_eG)$, is intimately related to the Lie bracket operation on T_eG. Let us describe it.

The map Ad. Consider a fixed element g of a Lie group G. The composition* of the left and right translations by g, $c_g := L_g R_{g^{-1}} : G \to G$, is the *inner automorphism* of G, $c_g : h \mapsto ghg^{-1}$. Since it is a diffeomorphism that preserves the point $e = c_g e$, its derivative at e is a linear invertible map of the tangent space to itself, $d_e c_g : T_e G \to T_e G$. This linear map, obtained from g, is denoted Ad_g (that is, $\mathrm{Ad}_g := d_e c_g$).

The map that associates the operator Ad_g to the group element $g \in G$, denoted $\mathrm{Ad} : G \to \mathrm{GL}(T_e G)$, is called *the adjoint representation of the group* G. It is in fact a representation of G on its Lie algebra $\mathfrak{g} = T_e G$:

$$\mathrm{Ad}_{gh} = \mathrm{Ad}_g \mathrm{Ad}_h.$$

The map ad. The differential of the map Ad, at $g = e$, is the linear map

$$\mathrm{ad} := d_e \mathrm{Ad} : T_e G \to \mathrm{End}\,(T_e G)$$

that associates to each vector $A \in T_e G$ the linear map $\mathrm{ad}_A : T_e G \to T_e G$,

$$\mathrm{ad}_A = \frac{d}{dt}\Big|_{t=0} \mathrm{Ad}_{g(t)}\,,$$

where $g(t)$ is any curve on G with velocity vector $\dot{g}(0) = A$ at $g(0) = e$.

The map ad is called *adjoint representation of the Lie algebra* $\mathfrak{g} = T_e G$.

Theorem 6. *The operator* ad_A *is given by the Lie bracket*: $\mathrm{ad}_A B = [A, B]$.

Proof. We shall apply the linear operator ad_A to a vector $B \in \mathfrak{g}$, for the case of the group of matrices $G = \mathrm{GL}(n, \mathbb{R})$.

For $h = 1 + \varepsilon B$ we get $c_g h = 1 + \varepsilon c_g B = 1 + \varepsilon g B g^{-1}$, and hence $\mathrm{Ad}_g B = g B g^{-1}$. Applying it along the curve $g(t) = 1 + tA$ we obtain

$$\mathrm{Ad}_{g(t)} B = (1 + tA) B (1 + tA)^{-1} = B + tAB - tBA + o(t)\,. \text{ So,}$$

$$\mathrm{ad}_A B = AB - BA = [A, B]. \tag{10}$$

Although there exist finite dimensional Lie groups which cannot be represented as Lie subgroups of the group of matrices $\mathrm{GL}(n, \mathbb{R})$, one similarly proves the expression $\mathrm{ad}_A B = [A, B]$ for any finite dimensional Lie group. $\square$

---

*The reader knows certainly that a linear map B of a vector space to itself is transformed into CBC^{-1} by the "linear change of variables" C.

"Non matricial" Lie groups. The double cover of $\mathrm{SL}(2,\mathbb{R})$, called *metaplectic group*, is a finite dimensional Lie group that is not a subgroup of $\mathrm{GL}(n,\mathbb{R})$ for any finite n. Another non matricial Lie group is obtained from the Heisenberg group H as the quotient $H_{\mathrm{Heis3}} = H/Z$, where

$$H = \left\{ \begin{pmatrix} 1 & a & c \\ 0 & 1 & b \\ 0 & 0 & 1 \end{pmatrix} \,\Big|\, a,b,c \in \mathbb{R} \right\} \quad \text{and} \quad Z = \left\{ \begin{pmatrix} 1 & 0 & n \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \,\Big|\, n \in \mathbb{Z} \right\}.$$

The general theory of abstract Lie groups was invented in the attempts of the algebraists to extend the results of the studies of the groups of matrices to some more general objects. However, the result of these difficult attempts was essentially negative in the sense that they have proved the absence of any new "more general" finite dimensional object: The non matricial Lie groups are obtained from the matricial ones by taking a cover —like the metaplectic group— or by taking a quotient by some discrete normal subgroup —like H_{Heis3} above. (Note that under these operations the Lie algebra is not changed.)
This absence of pathological finite dimensional generalisations is considered by some mathematicians as a great achievement of modern mathematics. But we prefer to consider it as a new confirmation of the old idea that the simple objects (like the groups of matrices and their subgroups) deserve more attention than the attempts to extend their theories. Consequently, we shall study the groups of matrices, like GL, SO, U.

7.8 Adjoint representation and Cartan subgroups

The great French mathematician Elie Cartan (student of Sophus Lie) continued the Lie theory of groups and introduced some special subgroups (called now Cartan subgroups to honour him). We shall see them below.

PROBLEM. Let A be a diagonal matrix with eigenvalues $(\lambda_1, \ldots, \lambda_n)$. Find the eigenvalues and the eigenvectors of the operator ad_A, which acts on the space $\mathbb{R}^{n^2} = \mathfrak{g}$ of arbitrary $n \times n$ matrices (sending B to $AB - BA$).

SOLUTION. The left multiplication of a matrix B by the diagonal matrix A multiplies the whole k-th line of B by the eigenvalue λ_k of A. The right multiplication by the diagonal matrix A multiplies the whole ℓ-th column of the given matrix B by the eigenvalue λ_ℓ of A.

Denote $B_{k,\ell}$ the matrix whose elements are all zero, except the only element equal to 1 staying at the intersection of the k-th line with the ℓ-th column. For these matrices we obtain the equality

$$\mathrm{ad}_A(B_{k,\ell}) = (\lambda_k - \lambda_\ell) B_{k,\ell}.$$

Thus these matrices $B_{k,\ell}$ $(1 \leqslant k \leqslant n,\ 1 \leqslant \ell \leqslant n)$ are the n^2 eigenvectors of $\mathrm{ad}_A : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$, with respective eigenvalues $\lambda_{k,\ell} = \lambda_k - \lambda_\ell$.

Remark. The eigenvalues $\lambda_{k,\ell}$ of ad_A depend linearly on the eigenvalues λ_j of the diagonal matrix A, but the *eigenvectors* $B_{k,\ell}$ of ad_A do not depend on the elements λ_j of A. This independence is natural because all diagonal matrices commute and, therefore, should have the same eigenvectors.

The non degenerate diagonal matrices (of diagonal linear operators in a fixed basis) form a commutative Lie subgroup of dimension n of the multiplicative group $\mathrm{GL}(n, \mathbb{R})$. It is called a *Cartan subgroup* of $\mathrm{GL}(n, \mathbb{R})$.

Of course, it depends on the coordinate system: a different coordinate system provides an isomorphic, but different, Cartan subgroup.

However, for the complex matrices (of linear complex operators) there exists an abstract version of the definition of the Cartan subgroups that avoids the explicit reference to any coordinate system.

Observe that the *Cartan subgroups* of complex diagonal matrices are the maximal commutative subgroups of the Lie group $G = \mathrm{GL}(n, \mathbb{C})$.

Then, for any complex Lie group one introduces the intrinsic

Definition. A *Cartan subgroup* of a complex Lie group G is a maximal commutative subgroup of G.

Problem. Find all Cartan subgroups of the Lie groups $\mathrm{GL}(n, \mathbb{C})$ and $\mathrm{U}(n)$.

Answer. Each Cartan subgroup consists of the non degenerate (respectively, unitary) diagonal complex matrices of order n in some complex (respectively, Hermitian orthogonal) basis of the complex vector space $\mathbb{C}^n$.

In these cases, given a Cartan subgroup H corresponding to a fixed basis, its Lie algebra is the vector space $\mathfrak{h}$ of diagonal matrices, and the eigenvectors and eigenvalues of the operator ad_A are provided by the same formulae as for the above case of the real diagonal matrices. Notice that the bracket operation restricted to $\mathfrak{h}$ is zero.

Remark. Since each eigenvalue $\lambda_{k,\ell}$ of the linear operator ad_A,

$$\lambda_{k,\ell} = \lambda_k - \lambda_\ell\,,$$

is a linear function of the eigenvalues of the diagonal matrix A, the linear functions $\lambda_{k,\ell} : \mathfrak{h} \to \mathbb{C}$, with $A \mapsto \lambda_{k,\ell}(A)$, can be considered as elements of the dual space of $\mathfrak{h}$,

$$\lambda_{k,\ell} \in \mathfrak{h}^*\,.$$

In the case of the unitary matrices, the vector space $\mathfrak{h}$ has a natural Hermitian structure, and hence the linear functions $\lambda_{k,\ell} : \mathfrak{h} \to \mathbb{C}$, are written as

$$\lambda_{k,\ell}(A) = \langle\, r_{k,\ell} \,, A \,\rangle \quad \text{for any} \quad A \in \mathfrak{h},$$

where $\{r_{k,\ell}\}$ is a remarkable "root*system" of vectors in Hermitian space $\mathfrak{h}$.

Namely, writing $(e_1, \ldots, e_n)$ for the natural Hermitian basis of $\mathfrak{h}$, which corresponds to the coordinate system $(\lambda_1, \ldots, \lambda_n)$, the "*root vectors*" are

$$r_{k,\ell} = e_k - e_\ell \,,$$

except the zero vectors $r_{k,k} = 0$ (obtained for $k = \ell$, with eigenvalue 0) which are not elements of the root system.

The system of 6 roots for the case U(3) is shown in Fig. 7.11. They are six vectors inside the plane of the traceless diagonal matrices $h_1 + h_2 + h_3 = 0$ (say, $h_1(r_{1,2}) = 1$, $h_2(r_{1,2}) = -1$, $h_3(r_{1,2}) = 0$).

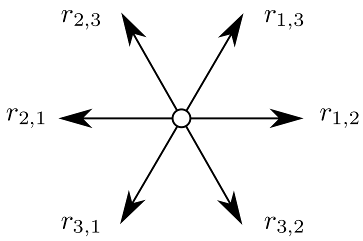


Figure 7.11: The root system of the Lie group U(3).

Similarly, the roots of the Lie group U(n) live in the $(n-1)$-dimensional plane $i\mathbb{R}^{n-1}$ of the traceless diagonal anti-Hermitian matrices (whose diagonal elements are purely imaginary, $h \in i\mathbb{R}^n$, and which have zero trace).

Exercise. Find the Cartan subgroups of the Lie groups SO(2), SO(3), SO(4), SO(n), and their root systems.

---

*The word "root" and the notation r reflect the origin of these vectors: They describe the eigenvalues, which are the roots of the characteristic equation. The reflection group generated by the reflections in mirrors orthogonal to the roots (that we shall study soon), generalises the monodromy group, which permutes the roots of the characteristic equation (over the complex paths of its coefficients in the complexified Lie algebra). All these objects were introduced to the Lie group theory by Killing, and consequently the monodromy group is usually called "Weyl group". The Weyl group of the unitary group U(n) is the symmetric group $S(n-1)$ formed by the symmetries of a simplex in the space $\mathbb{R}^{n-1}$.

7.9 Root Systems and Crystallographic Groups

The nice geometry of the root system of Fig. 7.11 is a particular case of a general theorem.

Reflections and Mirrors. A *Euclidean reflection* in a Euclidean vector space is an orthogonal transformation preserving one hyperplane point-wise, called the *mirror*, and reversing the normal line to this hyperplane.

Exercise. Consider a vector $v \in \mathbb{R}^n$ normal to the mirror of a Euclidean reflection. Denoting the inner product by $\langle \cdot, \cdot \rangle$, verify that such orthogonal transformation is given by

$$s_v : w \mapsto w - 2(\langle w, v\rangle / \langle v, v\rangle)v.$$

Root systems. In Euclidean vector space $\mathbb{R}^n$, a finite set R of non zero vectors that span $\mathbb{R}^n$ is called a *root system* if

(i) for each v in R the reflection s_v maps R into itself, and

(ii) for each pair of vectors v, w in R the map s_v sends the integer lattice spanned by v and w into itself. (Some authors replace this condition by the equivalent requirement that $2(\langle w, v\rangle / \langle v, v\rangle)$ be integer.)

The vectors of R are called *roots*. A root system is said to be *reduced* if for every root v the only colinear root to v is $-v$.

Weyl group. The *Weyl group* $W(R)$ of a root system R in Euclidean space $\mathbb{R}^n$ is the group generated by the Euclidean reflections in the mirrors orthogonal to the vectors of that root system.

Example. The Weyl group of U(3) (Fig. 7.11) contains 6 elements and is isomorphic to the group $S(3)$ of symmetries of an equilateral triangle.

Example. The Weyl group of the root system of the Lie group U(n) is finite. It is the symmetry group of the regular simplex in $\mathbb{R}^{n-1}$ (hypertetrahedron).

Example. The Weyl group of the root system of the Lie group SO(n) is also finite. It is the symmetry group of a (higher dimensional) cube if n is odd, while for even n it is a subgroup of the symmetry group of a cube (containing one half of these symmetries).

Crystallographic groups. A group of orthogonal reflections in Euclidean space $\mathbb{R}^N$ is said to be *crystallographic* if it preserves some sublattice isomorphic to $\mathbb{Z}^N$.

An "inverse" of the preceding discussion, is a theorem that associates a remarkable Lie group to any finite group generated by Euclidean reflections, provided that such group be crystallographic.

In the case of the root system of Fig. 7.11 the lattice $\mathbb{Z}^2$ is generated by the roots (see Fig. 7.12): it suffices to take the basis $r_{1,2} = e_1 - e_2$, $r_{2,3} = e_2 - e_3$ of the plane $h_1 + h_2 + h_3 = 0$ of $\mathfrak{h} = \mathbb{R}^3$.

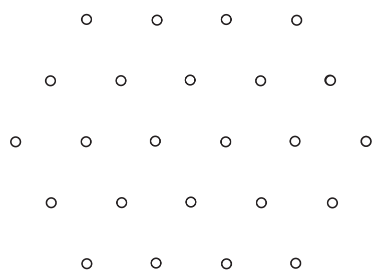


Figure 7.12: The crystallographic lattice of the root system of the group U(3).

The reflections in a finite set of mirrors may also generate an infinite group.

Problem. Find all the finite groups generated by reflections in Euclidean 2-dimensional vector space $\mathbb{R}^2$.

Answer. They are the symmetry groups of the regular polygons.

Solution. Two mirrors in the plane define a finite reflection group if and only if the angle between them is commensurable with 2π. This reflection group is then the symmetry group of a regular p-gone, denoted by $I_2(p)$.

Problem. Which of the finite groups generated by reflections in Euclidean 2-dimensional vector space $\mathbb{R}^2$ are crystallographic?

Answer. The group of symmetries of a regular k-gon preserves some crystallographic lattice (isomorphic to $\mathbb{Z}^2$) if and only if $k = 2, 3, 4$ or 6 – Fig. 7.13.

A crystallographic group (in a Euclidean space) has no elements of order 5, nor elements whose order is different from the above four numbers.

Thus, the symmetry groups of the regular pentagons, heptagons, octagons, and so on are not crystallographic, though they are related to the quasi-crystals of physicists. Whenever you see a crystal with a 5-th order symmetry, it is fake.

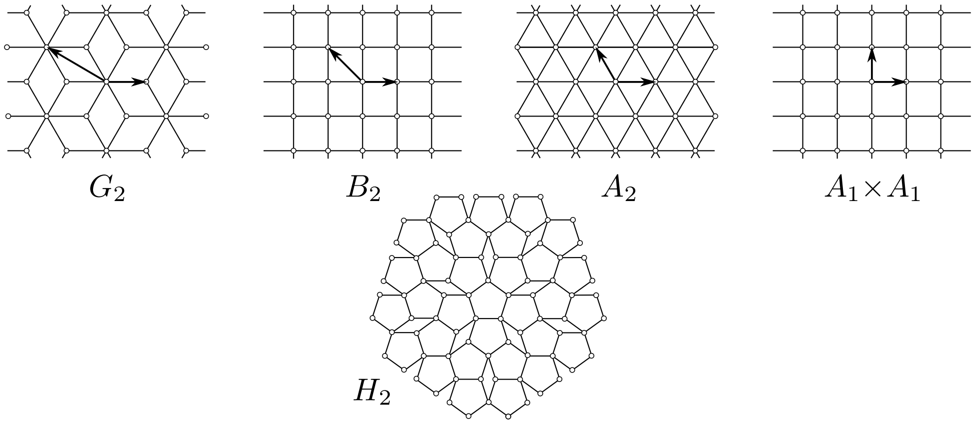


Figure 7.13: Lattices preserved by crystallographic plane reflection groups (together with the orthogonal vectors to the generating mirrors of these groups) and a non crystallographic reflection group.

There exists however some *quasi-crystalline* structures which may have symmetries of any order and whose general theory is explained in the books [24] and [22]. We can see an example in Fig. 7.14.

7.10 Reflection Groups and Dynkin Diagrams

The classification of Euclidean reflection groups (that is, of the finite groups generated by reflections in Euclidean vector spaces) is due to H. Coxeter.

The generating reflections are usually described by the so-called Dynkin diagrams (invented by Killing, who used them to classify the Lie groups). Namely, the Dynkin diagram of a reflection group provides a short description of that group in terms of the angles between the generating mirrors. Such a diagram is a graph whose vertices represent the orthogonal vectors to the mirrors, and whose edges represent the angles between these vectors.

Namely, two vertices are not joined if they represent orthogonal vectors; they are joined by a simple edge if the angle is 120°, while a double edge and a triple edge represent the respective angles of 135° and of 150°.

Example. The diagram $A_2 = \circ\!\!-\!\!\!-\!\!\circ$ represents two mirrors forming an angle of 120 degrees in the plane. The corresponding reflection group is the symmetry group of an equilateral triangle.

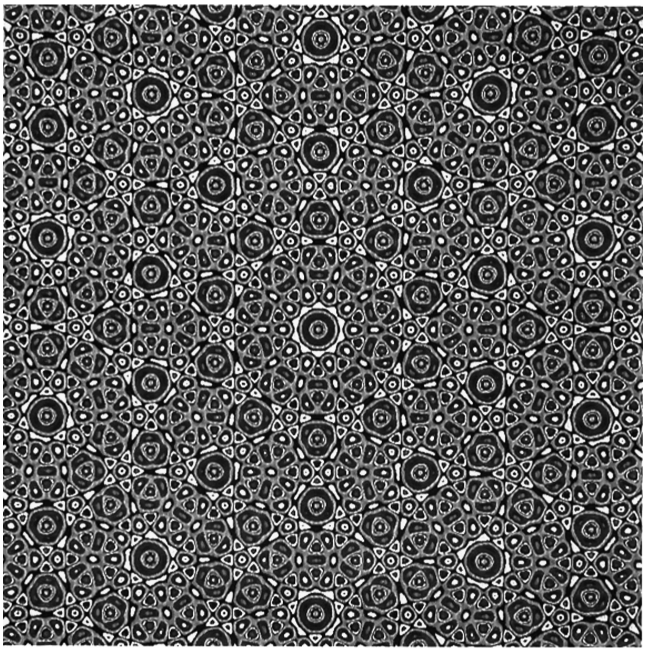
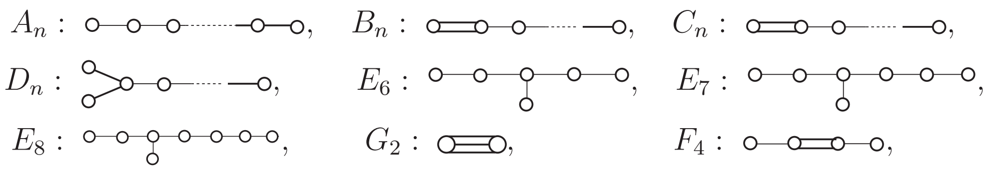


Figure 7.14: A quasi-crystalline function $f(x) = \sum \cos(v_k, x)$ having a 5 order symmetry, for the 5 vertices v_k of a regular pentagon.

The *irreducible reflection groups* are those which have no subspaces invariant under the action of the whole group. The orthogonal direct sum of two Euclidean spaces equipped with reflection groups, is a Euclidean space on which the direct product of those groups acts naturally as a new (reducible) reflection group. Any reducible reflection group is isomorphic to such an orthogonal sum of irreducible components acting on invariant subspaces.

Coxeter's classification separates the finite irreducible Euclidean reflection groups into *crystallographic* and *non-crystallographic groups*:

1. List of irreducible crystallographic reflection groups. It consists of four infinite series (A, B, C, D) and five exceptional cases $(E_6, E_7, E_8, G_2, F_4)$:

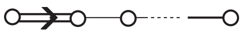


where the subindex means the dimension of Euclidean space and the number of vertices of the diagram.

The groups B_n and C_n are isomorphic, but there are two different lattices preserved by these groups, and the generating vectors of these lattices have different lengths in the crystals B_n and C_n. Thus, to distinguish their diagrams we draw an arrow pointing from the longer to the shorter root: The symbol  means that the left vertex represents a vector whose length is $\sqrt{2}$ times bigger than that of the vector represented by the right vertex.

In this notation the corresponding lattices are described by the diagrams

B_n :  (many longer vectors),

C_n :  (many shorter vectors),

F_4 : .

In the case of the group G_2 the lengths of the two vectors are also different : .

Remark (Crystallographic groups $\leftrightarrow$ Lie algebras). The list of irreducible crystallographic groups (considered with their lattices) coincides with the list of Weyl groups of the root systems of the simple (complex) Lie algebras.

2. List of non-crystallographic irreducible reflection groups. It consists of the series $I_2(p)$, $p = 5, 7, 8, 9, \ldots$, (that is, $p \neq 2, 3, 4, 6$) and of the two sporadic groups H_3, H_4, where H_3 is the icosahedron symmetry group (with 120 elements) and H_4 is the "hypericosahedron" symmetry group (with 120^2 elements)* that we shall now describe.

Hypericosahedron. The hypericosahedron is a regular convex polyhedron with 120 vertices in Euclidean space $\mathbb{R}^4$. To construct it, we consider, among the icosahedron symmetries, the subgroup Γ of the 60 icosahedron rotations. This group Γ is a subgroup of the group of rotations of Euclidean space $\mathbb{R}^3$, $\Gamma \subset \mathrm{SO}(3)$. The double (universal) covering $\mathrm{Spin}(3) \to \mathrm{SO}(3)$ (see p. 102) covers the subgroup Γ of the rotation group SO(3) by the *binary icosahedron group* $\widehat{\Gamma} \subset \mathrm{Spin}(3) = \mathrm{SU}(2)$ which consists of 120 elements.

But the group $\mathrm{Spin}(3) = \mathrm{SU}(2)$ is the group $\mathbb{S}^3$ of quaternions of norm 1 (see p. 94). Thus we have constructed a set $\widehat{\Gamma} \subset \mathbb{S}^3$ of 120 points on the unit sphere of Euclidean space $\mathbb{R}^4$.

These 120 points are the vertices of a regular 4-dimensional polyhedron which is their convex hull and is called the *hypericosahedron.* It has 600

---

*We shall not prove that there are no other non-crystallographic irreducible reflection groups than those listed above, but Coxeter proved it.

3-dimensional tetrahedral faces. Its Euclidean symmetry group, denoted by H_4, is a group of reflections which acts on $\mathbb{R}^4$ and which is obtained from the binary group in the following way.

The binary group $\widehat{\Gamma}$ acts on itself by the left rotations (L_g) and the right rotations (R_g), which commute and define, hence, the isometric action of the direct product $\widehat{\Gamma} \times \widehat{\Gamma}$ on $\widehat{\Gamma}$, and hence on $\mathbb{R}^4$.

These 120^2 symmetries form the group H_4 of all Euclidean symmetries of the hypericosahedron.

Remark. Although Bourbaki's book on root systems has been very fruitful in representation theory and is certainly their best book, this Coxeter theory of the group H_4 was perhaps too geometric for them that they have replaced it by the tables of the 4×4 matrices of the elements of the group H_4. Anyway, nobody is safe of lacking a simple description of an object.

7.11 On Logical Reduction of Axioms and General Theories

The classification of the crystallographic groups generated by reflections in Euclidean spaces, provides a general theory of special Lie groups (which contains the unitary groups, the orthogonal groups and the symplectic groups). These Lie groups are essentially reducible to the groups of matrices (where the calculations are easier, and where everything had been discovered earlier). The groups whose study is reducible to the elementary case of the matrix groups, are called *classical groups*, going back to Weyl.

Developing the general theory of classical groups, one has to avoid to mention the real origins of the notions and of the theorems: one should avoid "matrices", "diagonal matrices", "coordinate changes", using instead "Cartan subgroups" and "adjoint representations".

The gain is the possibility to prove similar facts for different groups only once. So, replacing the words, "the case of the orthogonal matrices is similar to that of the general non degenerate matrices" by a formal proof that provides the generalised versions containing the original ones.

Essentially, the achievement is the paper economy: one would need no new ideas to deal with the more general classical groups than with the groups of matrices. However, the history of sciences shows that similar paper economy reasons produced a lot of important mathematical works. For example, Euclid's book.

All Euclidean geometry had been discovered and written millennia earlier than Euclid's book (including, say, the Pythagoras theorem and its proof, written by the Khaldeans and known to the Egyptians, who had explained it to Pythagoras).

Euclid, wanting to achieve the paper economy, reduced the set of axioms of the Egyptian geometry. Namely, he chose his famous "fifth postulate" (claiming that *through a point outside a straight line in the plane there passes one and only one line parallel to the original one*) from a set of several axioms of the Egyptian geometry (including, for instance, the property of the sum of the angles of a triangle).

Observing that some of these axioms might be deduced from the other axioms, Euclid made theorems from some of the Egyptian axioms, shortening in this way his text-book.

The main mathematical result of this novelty is the proof of the *dependence of the old axioms*: no new properties of the real world were discovered.

However, all this old Egyptian geometry is called now "Euclidean", and it contains "Pythagoras theorem", which had been a state secret of Egyptians and which was never published by Pythagoras, who had promised in Egypt never divulge this secret.

Nowadays, the safest way to make a scientific career, specially in mathematics, is to create generalisations of known things and to publish them under your name, forgetting the references to the original discoverers (America is not Columbia). Professor M. Berry stated the "Arnold eponymic principle": *Whenever an object or a theory has a personal name, it is not the name of the discoverer.* But he added an important remark, called by him "Berry's principle": *Arnold's principle is applied to Arnold's principle.*

We shall not continue the discussion on the theory of the classical groups here: There is a lot of good text-books on the subject, say, that of W. Fulton and J. Harris [73]. Instead, we shall add some remarks on the problem of the independence of the axioms.

The great *thesis of Lobachevsky*, announced by him in the XIXth century, was the *independence* of the 5-th Euclid postulate, from the other axioms of Euclidean geometry.

He had tried to prove this postulate by assuming that the opposite postulate takes place, and then trying to find a contradiction in the theory.

Lobachevsky postulate. *For a point outside a straight line in the plane there passe several straight lines having no common point with the original straight line.*

However, he was unable to find any contradiction in the resulting theory, called today "Lobachevsky geometry". This inability does not prove his main thesis on the independence of the Euclidean axioms: it might be explained by the weakness of the attempts of Lobachevsky. But insisting on his *independence thesis* he created "Lobachevsky geometry", though without achieving his goal. Today this goal is achieved.

In Chapter 8, we shall give the proof of the main Lobachevsky thesis.

7.11.1 Appendix on the Tensor Product

Definition. The *tensor product* $X \otimes Y$ of two real vector spaces* X and Y is the real vector space formed by all the bilinear forms (defined) on the direct product $X^* \times Y^*$ of the dual spaces X^* and Y^* (i.e., of the spaces of linear forms in X and in Y).

*Real vector spaces are, in algebraic abstract terms, the $\mathbb{R}$-modules, and the tensor products may be defined for the modules over other rings of algebras, but we shall not repeat the (obviously similar) definitions in these more general cases (see, e.g. Eisenbud's book [61])

Thus, denoting by $\mathrm{bi}(A, B)$ the vector space of the bilinear forms of arguments $a \in A$ and $b \in B$, we can write

$$X \otimes Y = \mathrm{bi}(X^*, Y^*).$$

If $\dim X = k$ and $\dim Y = \ell$, then $\dim X^* = k$ and $\dim Y^* = \ell$ and therefore $\dim(X \otimes Y) = k\ell$. Moreover, each pair $x \in X$ and $y \in Y$ determines an element of $X \otimes Y$, denoted by $x \otimes y$: it is the bilinear form whose value on the pair $a \in X^*$, $b \in Y^*$ is

$$(x \otimes y)(a, b) = a(x)b(y).$$

Exercise. Verify that the following three properties hold

$$(cx) \otimes y = x \otimes (cy) = c(x \otimes y) \quad \text{for any } c \in \mathbb{R},\ x \in X,\ y \in Y;$$
$$(x_1 + x_2) \otimes y = (x_1 \otimes y) + (x_2 \otimes y) \quad \text{for any } x_1, x_2 \in X,\ y \in Y;$$
$$x \otimes (y_1 + y_2) = (x \otimes y_1) + (x \otimes y_2) \quad \text{for any } x \in X,\ y_1, y_2 \in Y.$$

These evident properties could be taken as the definition of the tensor product, whenever one would like to enhance the authority of trivial algebra, that is, of the theory of notations.

In the same way, the multiple tensor products, say of three vector spaces, may be defined as the vector space of the trilinear forms on dual spaces, and similarly for tensor products of more vector spaces. This “multiplication” of vector spaces is an associative operation: $(X \otimes Y) \otimes Z = X \otimes (Y \otimes Z)$.

Symmetric and anti-symmetric tensor products. In the case of equal factors, like in $X \otimes X$, one can consider the spaces of the symmetric forms or of the anti-symmetric forms (of dimensions $k(k+1)/2$ and $k(k-1)/2$, respectively, if $\dim X = k$). These vector spaces are called respectively the *symmetric* and *anti-symmetric tensor product spaces*. The later one is also called the *exterior product space* and is usually denoted by $X \wedge X$, and one has $x_1 \wedge x_2 = -x_2 \wedge x_1$ (similarly, the anti-symmetric product of m copies of a vector space is denoted by $X \wedge X \wedge \ldots \wedge X = \wedge^m X$ and it is also called exterior product).

Example. The vector space of the exterior m-forms on a real vector space V of dimension n is the exterior product of m copies of the dual space: $\wedge^m V^*$. It consists of the anti-symmetric m-linear forms (on m elements $x_1 \in (V^*)^* = V, x_2 \in V, \ldots, x_m \in V$).

This notation was already used above, and now we have explained its origin.

Example. The m-th exterior product $\wedge^m V$ consists of the so-called m-vectors in V: An *m-vector* in V is a sequence of m vectors of V. When two such sequences form the same set, they are considered as the same m-vector if the permutation transforming one sequence to the other is even. If it is odd, then they are considered as two opposite elements of the vector space of m-vectors.

One defines the linear combinations of m-vectors by imposing the linearity with respect to the first vector

$$(x_1 + x_1', x_2, \ldots, x_m) = (x_1, x_2, \ldots, x_m) + (x_1', x_2, \ldots, x_m),$$

and similarly with respect to each vector x_k $(1 \leqslant k \leqslant m)$. This real vector space, formed by the m-vectors of the n-dimensional vector space V, has dimension C_m^n, and it is naturally dual to the space of the m-forms $\wedge^m V^*$.

The value of an m-form ω on an elementary m-vector is provided by the definition of the m-form. The value of the form on a linear combination of elementary m-vectors is defined as the linear combination of the values on the elementary summands.

In this sense the algebra of forms represents some algebraic dual version of the geometry of multi-vectors. Up now, we have avoided this dual point of view studying the values of the forms on the elementary multi-vectors rather than on their linear combinations.

In most cases, we shall continue this geometric study of the exterior algebra of forms. Elementary multi-vectors (also called decomposable multi-vectors) are for the general multi-vectors what are the monomials for the polynomials.

The symmetric tensor products are also useful in many problems.

Example. A Riemannian metric on a manifold M is a field of positive definite quadratic forms on its tangent spaces.

These quadratic forms belong to the symmetric tensor product of the cotangent spaces, $X \otimes_S X$, $X = T_x^* M$.

A pseudo Riemannian metric is defined similarly, relaxing the positive definiteness condition. The special cases of forms of signatures of the types $\pm(x_0^2 - x_1^2 - x_2^2 - \ldots - x_n^2)$ are called *relativistic Lorenzian metrics.*

The second (quadratic) fundamental forms of the co-oriented hypersurfaces in Euclidean spaces (defined in Ch. 10, p. 315) are also elements of the symmetric tensor products of the cotangent spaces of the hypersurfaces.

Chapter 8

Lobachevsky geometry

In order to do some "geometry" on manifolds, analogous to Euclidean geometry, one needs to consider lengths, areas, angles, etc. of geometrical figures in those manifolds. For a manifold M embedded in a Euclidean space $\mathbb{R}^n$ one uses the metric on this ambient space to measure the lengths of curves in M, angles between vectors tangent to M, volumes, etc. One express all these quantities by means of the lengths of tangent vectors, that is, by means of their Euclidean norm, which is given by the positive-definite quadratic form defining the Euclidean metric in $\mathbb{R}^n$.

There are however several situations and problems in which the geometrical context leads naturally to measure lengths of tangent vectors using a metric given by a quadratic form that depends smoothly on the points of the considered manifold, with no relation to any metric of any ambient Euclidean space. When such a metric (positive-definite quadratic form) is given at each tangent plane of a manifold M, one says that M is a *Riemannian manifold*. For example, the velocity of light in the atmosphere depends on the air density, which is essentially a function of the height. Thus the velocity vector of a light ray is measured with a "Riemannian metric" that depends on the position at which we are measuring (for this reason light rays in atmosphere are curvilinear).

In this chapter we study the geometry of the Lobachevsky plane as a first example of a Riemannian 2-dimensional manifold. A more general study of Riemannian manifolds of any dimension is given in Chapter 9.

8.1 Poincaré model of Lobachevsky plane

Consider the group G of orientation-preserving affine maps of the affine real line $\mathbb{R} = \{x\}$. An element g of the group G is an affine transformation $g : \mathbb{R} \to \mathbb{R}$, defined by the formula $g(x) = ax + b$, $a > 0$. Thus, the Lie group G is the upper half-plane, $\mathbb{R}^+ + \mathbb{R} = \{a, b : a > 0\}$ (Fig. 8.1).

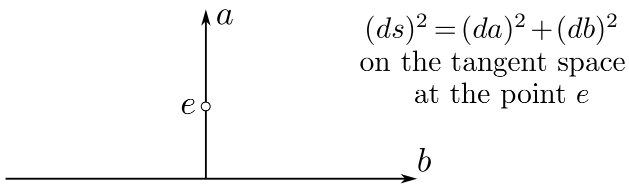


Figure 8.1: Lobachevsky plane $\{a, b : a > 0\}$.

The product $g \cdot g'$, where $g'(x) = a'x + b'$, sends x to

$$a(a'x + b') + b = (aa')x + (b + ab').$$

Fix some positive definite quadratic form on the tangent space $\mathfrak{g} = T_eG$, where the identity transformation $e \in G$ is the point $(a = 1, b = 0)$ of the half-plane. The left translation L_g identifies the tangent space T_eG with the tangent space T_gG. Translate the chosen quadratic form to all the points g by the left translations. Thus a left-invariant field of positive definite quadratic forms is defined on tangent spaces of G, that is, a (left invariant) Riemannian metric on the manifold G.

Definition. A *Riemannian manifold* is smooth manifold with a prescribed positive-definite quadratic form on every tangent space T_xM. The quadratic form is called *Riemannian metric*.

Definition. The *Lobachevsky plane* is the group G equipped with the left invariant Riemannian metric $(ds)^2$ (represented at e by the quadratic form $(ds)^2 = (da)^2 + (db)^2$).

Theorem 1. *The Riemannian metric of the Lobachevsky plane is defined by the quadratic form*

$$(ds)^2 = \frac{(da)^2 + (db)^2}{a^2} \ . \tag{1}$$

Proof. The left translation operator L_g, obtained from the multiplication by $g : x \mapsto ax + b$, sends the motion $a'(t) = 1 + \alpha t$, $b'(t) = \beta t$ (that starts at time $t = 0$ from e with velocity (α, β) in our coordinates a, b), to the motion

$$a(t) = a(1 + \alpha t) \ , \quad b(t) = b + a\beta t \ ,$$

according to the multiplication rule in the group G, written above. At $t = 0$, this translated motion starts at the point a with velocity $(\dot{a} = a\alpha, \ \dot{b} = a\beta)$.

Therefore, $\alpha = \dot{a}/a$, $\beta = \dot{b}/a$ provide any velocity vector $(\dot{a}, \dot{b}) \in T_g G$, and the left invariant form $(ds)^2$ takes on it the value

$$\alpha^2 + \beta^2 = \big((\dot{a})^2 + (\dot{b})^2\big)/a^2. \qquad \square$$

If we were started from any positive definite quadratic form at e, we would arrive to an isometric Riemannian manifold G, up to a constant factor in the metric (similar to the difference of the spherical metrics on the spheres of different radii).

PROBLEM. If we transport the metric by right translations instead of left translations, we shall obtain a different Riemannian metric on G. But the manifolds G with the right invariant and with the left invariant metrics are isometric. Find the isometric map.

ANSWER. The group inversion, $g \mapsto g^{-1}$, sends a left invariant metric to a right invariant one.

The upper half-plane, equipped with the Riemannian metric (1), is called *the Poincaré model of the Lobachevsky plane* (in spite of the fact that Poincaré was the first to propose it).

8.2 Geodesics of Lobachevsky plane and Euclide's 5th postulate

We shall study now the properties of this geometrical object, proving, among other things, that it verifies the Lobachevsky axiom: Given a point and a straight line not containing it, there exist many straight lines containing the given point and not intersecting the given line.

Lemma 1. *All the shifts $(a, b) \mapsto (a, b+c)$, with $c \in \mathbb{R}$, preserve the distances in the Lobachevsky plane* (the shifts are Lobachevsky isometries).

Proof. $d(b+c) = db$, hence ds^2 is preserved (the shift is also a left translation). $\square$

Lemma 2. *Any homothety $(a, b) \mapsto (ca, cb)$, with $c > 0$, is an isometry of the Lobachevsky plane.*

Proof. The fraction (1) is preserved, since the functions da, db and a are multiplied by the same number c. $\square$

Lemma 3. *The inversion with respect the unit circle* $a^2 + b^2 = 1$ (sending a point on a ray through the origin at distance r from it to the point on the same ray at distance $1/r$ from the origin) *is an isometry of the Lobachevsky plane.*

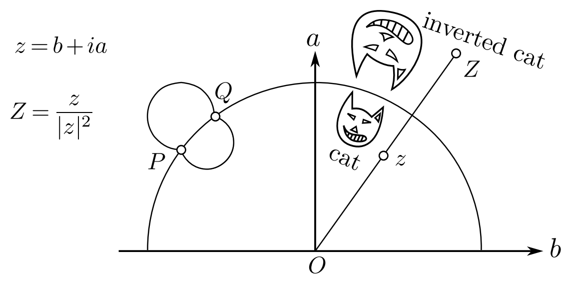


Figure 8.2: The inversion isometry of the Poincaré model of the Lobachevsky plane (the reflection in the circle $a^2 + b^2 = 1$).

Proof. (Fig. 8.2). Denote the complex number $b + ia$ by z. In this notation the inversion sends z to $Z = 1/\bar{z} = z/|z|^2$. Therefore, the vectors of the tangent space T_zG are sent to T_ZG by the conjugate to the complex derivative,

$$dZ = \overline{(-1/z^2)\ dz}\ .$$

The Euclidean lengths of the tangent vectors are therefore divided by $|z|^2 = r^2$. The ordinate a/r^2 of the point $Z = z/|z|^2$ is also r^2 times smaller than the ordinate a of the point z. Therefore, the Lobachevsky plane length of the tangent vector at the point z is equal to the Poincaré model length of its image at the point Z, and therefore the inversion map is an isometry. □

Corollary. *The circle* $a^2 + b^2 = 1$ *is a geodesic line of the Poincaré model of the Lobachevsky plane.*

Proof. (Fig. 8.2). If the shortest line PQ, connecting two neighbouring points P and Q of the circle, were different from the circle, its reflected version would be a different line of the same length, connecting them. But a general theorem of calculus of variations claims that there is only one shortest path, provided that the points are sufficiently close to each other. The contradiction proves that in the Poincaré model of the Lobachevsky plane, the shortest path is circular. □

PROBLEM. Prove that *any semicircle orthogonal to the border line* $a = 0$ *of the half-plane* G (that is, whose centre lies on this border line) *is a geodesic of the Poincaré model of the Lobachevsky plane.*

Hint. This case can be reduced to the case of Lemma 3 by the isometric transformations of Lemma 1 and of Lemma 2.

Remark. The geodesics of the metric (1) may be also deduced from the kinematics of a free point in Lobachevsky plane (no force acting on it and hence no potential energy). To compute the orbit of a free point from Hamilton equations of motion, we use that the Lagrangian function is just the kinetic energy (since there is no potential energy):

$$L = \frac{1}{2}\frac{\dot{a}^2 + \dot{b}^2}{a^2}\,.$$

It provides the momenta $p_a = \dot{a}/a^2$, $p_b = \dot{b}/a^2$. The Hamilton function, obtained as the Legendre transform of L (see Section 16.7.2.), is equal to

$$H = a^2\left(\frac{p_a^2}{2} + \frac{p_b^2}{2}\right)\,.$$

Since b does not appear in H, Hamilton equations for p_b and b, $\dot{p}_b = -\frac{\partial H}{\partial b}$ and $\dot{b} = \frac{\partial H}{\partial p_b}$, imply the preservation law of the momentum $p_b = \varkappa = \text{const}$ and the equality $\dot{b} = \varkappa a^2$.

Taking into account the preservation law of the energy (here $E = L$) we have

$$\dot{a}^2 + \dot{b}^2 = 2Ea^2 \quad \text{and} \quad \dot{b} = \varkappa a^2\,.$$

From these equalities we get $\dot{a}^2 = 2Ea^2 - \varkappa^2 a^4$ and, hence,

$$\frac{da}{dt} = a\sqrt{2E - \varkappa^2 a^2}\,,$$

which leads to the circular orbits by an elementary integration.

In this algebraic way we eliminate the reference to the variational calculus theorem on the uniqueness of the geodesic, but we prefer the above geometrical proof based on the geometric Lemma 3.

The momentum preservation $p_b = \varkappa$ is essentially the Snell law in an optical medium that is homogeneous along the horizontal layers $a = \text{const}$. Compare with the geometric presentation of Lobachevsky plane as optical medium given in Section 8.6.

Lemma 4. *The straight half-lines* $b = \text{const}$, *orthogonal to the bordering line* $a = 0$ *of the half-plane* G, *are geodesic curves of the Poincaré model of the Lobachevsky plane.*

Proof. The orthogonal reflection in such a line is an isometry that preserves that line. In consequence, it is a geodesic line, according to the uniqueness of the shortest path (like in the proof of the above corollary). $\square$

Remark. Of course, these lines also describe some solutions of the above Hamilton equations, which provide another proof of Lemma 4, independent of the general theorems of calculus of variations.

The bordering line $a = 0$ of the Poincaré model is called the *absolute*. It does not belong to the half-plane G. The distance from e to the absolute is infinite, since the integral $\int_0^1 ds = \int_0^1 \frac{da}{a}$ diverges at $a = 0$.

Lemma 5. *All the geodesic lines of the Poincaré model of Lobachevsky plane are the Euclidean circles and the Euclidean straight lines that are orthogonal to the absolute* (described in the preceding problem and in Lemma 4).

Proof. For any direction at any given point of G there is either a circle or a straight line, orthogonal to the absolute, containing that given point and having there the prescribed direction (Fig. 8.3).

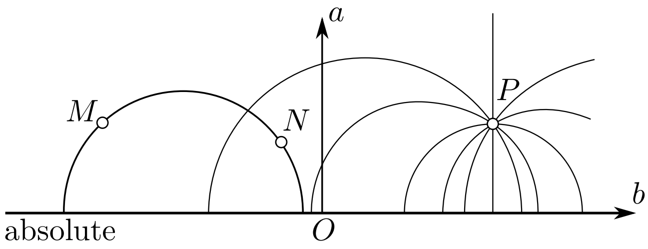


Figure 8.3: Some straight lines containing the point P, in the Poincaré model of Lobachevsky plane; and the straight line connecting the points M and N.

But at each point there is only one geodesic with a given direction (this is the uniqueness theorem of the theory of the ordinary differential equations). In consequence, there are no other geodesics in the Poincaré model. □

Theorem 2. *The Lobachevsky axiom is fulfilled in the Lobachevsky plane, and the Euclidean* 5*-th postulate is violated there.*

Proof. Take a point P exterior to the straight line $b = 0$ (to the a axis in Fig. 8.3). Clearly, there are many Lobachevsky straight lines (geodesics) that contain the point P and do not intersect the straight line $b = 0$. Similar examples are easily constructed for any point P and any Lobachevsky straight line, provided that it does not contain P. □

Remark. All the other axioms of the Euclidean plane geometry hold also in the Lobachevsky plane. For example, there is one and only one straight line, connecting any two points M and N of G (prove it, finding the centre on the absolute), and so on.

By the way, the *angle* between two curves in the Poincaré model (that is, between their tangent vectors at the intersection point) is equal to the Euclidean angle between them.

Indeed, in formula (1) the division of the form $(da)^2+(db)^2$ by the constant a^2 does not change the angles. However, the "orthogonality to the absolute line" in Lemmas 3–5 refers to the Euclidean angles, since the absolute line is far at infinity (and is not included in the Lobachevsky plane).

Theorem 3. *The Euclid postulate of the parallel lines is independent from the others axioms of the Euclidean geometry: It can not be deduced from the other axioms (provided that the Euclidean geometry itself has no interior contradiction).*

Proof. If the Euclidean axiom of the parallels were deducible from the other axioms, this deduction would remain true in the Poincaré model of the Lobachevsky plane (where the other axioms are fulfilled).

However the Euclidean axiom of parallel lines is not true in the Poincaré model (see the preceding theorem).

Thus, the deduction of this axiom from the others would imply a contradiction in the *Euclidean* geometry of the straight lines and the circles orthogonal to the absolute. Consequently, the axiom of the parallels cannot be proved from the other axioms (provided that Euclidean plane geometry is free of contradictions). □

We see now that the main Lobachevsky statement (on the independence of the 5-th postulate) was true and is now proved (in spite of the inability of Lobachevsky to prove it).

This is a remarkable example of an impossibility result in mathematics (similar to the impossibility of the rationality of $\sqrt{2}$, or of the construction of the centre of a circle with a ruler, having no compass, or of to find a formula in terms of radicals for the roots of the equation $x^5 + ax + 1 = 0$).

The incommensurability of the diagonal with the side of a square was kept as a secret by Pythagoras, because of his fair that the society and the state would no longer support mathematicians, knowing that their science (arithmetics of rational numbers) is insufficient even for the measuring of the lengths of real world objects (like the square's diagonals).

The unsolvability of the degree 5 equations in expressions containing radicals, proved by Abel, was hidden by Cauchy (perhaps being jealous to the success of the unknown young Norwegian).

A celebrated impossibility result is the Cohen solution (1963) of the "continuum problem" of Cantor: Does there exists an infinite subset M of the real line segment I that can not be mapped bijectively neither to I nor to $\mathbb{Z}$? The answer is the proof of the impossibility of a solution, similar to the impossibility of proving the 5-th postulate of Euclid. Thus, similarly to the coexistence of Euclidean plane and of Lobachevsky plane geometries, there exist two different mathematics: In one of them the Cantor subset M does exist, while in the other there exist no such subsets.

The proof is similar to the above proof of the Lobachevsky thesis : There is a model of the first mathematics in the second and a model of the second in the first (which contradicts both the positive and the negative answers to Cantor's question, provided that one of the two mathematics is free of contradictions).

Gödel's theorem implies that any formal mathematical theory (sufficiently deep to contain the elementary arithmetics of $2 + 2 = 4$) contains statements that cannot be neither proved nor disproved inside that theory.

We shall prove now *the impossibility of constructing the centre of a circle by means of ruler* (without compass).

Consider a cone constructed over a circular disc of a plane, such that its vertex does not lie on the straight line orthogonal to the disc at its centre.

Any ruler construction in the plane of the circle can be projected (by the rays connecting the geometrical objects to the cone vertex) to any other plane not containing the vertex. Among such planes there are some planes (Fig. 8.4) whose sections by the cone are circular: the initial plane with its parallel planes and some different planes (inclined similarly to the cone, but in the opposite direction).

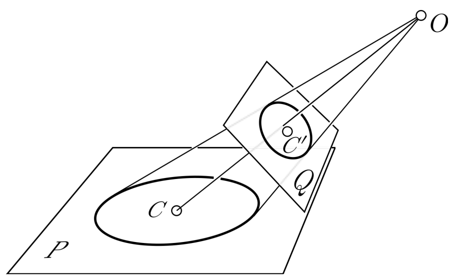


Figure 8.4: Projection from plane P to plane Q from the vertex O of the cone.

The whole system of straight lines that construct the centre of the initial circular disc in the plane P, will be projected to a similar system of straight lines in the plane Q. If this construction were lead always to the centre, we would obtain in Q the centre of the circle along which Q intersects the cone, on the line OC connecting O with the centre of the initial disc. However, the centre C' of this new disc does not belong to the line OC (as easy calculations show). Therefore the straight lines construction, leading always to the centre, is impossible.

8.3 The Geometry of Lobachevsky plane

We shall prove some remarkable facts relating the Lobachevsky geometry to many domains of mathematics and physics.

Theorem 4. *Let* $\left(\begin{smallmatrix}\alpha & \beta\\ \gamma & \delta\end{smallmatrix}\right)$ *be a real matrix of determinant one* ($\alpha\delta - \beta\gamma = 1$). *The complex projective map*

$$z \mapsto \frac{\alpha z + \beta}{\gamma z + \delta}$$

defines an isometry of the Poincaré model of the Lobachevsky plane G.

Proof. The real line $a = 0$ (of the complex numbers $z = b + ia$ of the plane $\mathbb{C}$) is sent to itself, since the matrix is real.

The point e (where $z = i$) is sent to the point

$$\frac{\alpha i + \beta}{\gamma i + \delta} = \frac{(\alpha i + \beta)(-\gamma i + \delta)}{\gamma^2 + \delta^2} = \frac{(\beta\delta + \alpha\gamma) + i(\alpha\delta - \beta\gamma)}{\gamma^2 + \delta^2}$$

of the upper half-plane, since $\alpha\delta - \beta\gamma > 0$.

Therefore, the upper half-plane is sent onto itself.

To see that our map is an isometry, it suffices to check that it is a product of shifts (of Lemma 1), of homotheties (of Lemma 2) and of inversions (of Lemma 3), the number of inversions should be even, to preserve the orientation of G:

$$z \mapsto \left(\frac{C}{\left(A + \left(D\left(\frac{1}{z}\right)\right)\right)}\right) + B = \frac{Cz + B(D + Az)}{D + Az} = \frac{\alpha z + \beta}{\gamma z + \delta}\ ,$$

for $A = \gamma$, $B = \beta/\delta$, $C = 1/\delta$ and $D = \delta$ if $\delta \neq 0$, with easier formulas for $\delta = 0$ (the brackets in the first term indicate the successive isometries). $\square$

Remark. The complex projective maps of the preceding theorem preserve the metric and the orientation of the Poincaré model. Hence the group $\mathrm{SL}(2, \mathbb{R})$ of the real matrices of determinant 1 acts on the Lobachevsky plane as its "rotation group" (of the preserving orientation symmetries).

Consider the straight lines of the Lobachevsky plane, intersecting the absolute circle at the same point in the Poincaré model (to which we include the point $z = \infty$ of $\mathbb{C}\mathrm{P}^1$).

In the model they form a 1-parameter family of Euclidean circles orthogonal to the absolute straight line at the same point. Choosing this point to be $z = \infty$, we get the family of the "vertical lines", $b = \text{const}$.

In the Poincaré model of the Lobachevsky plane, the angles between the intersecting curves coincide with the angles in the sense of the Euclidean plane geometry. Thus, in the Poincaré model the curves orthogonal to the "vertical lines" $b = \text{const}$ are the horizontal lines $a = \text{const}$. These lines are straight lines in Euclidean geometry, but in the Lobachevsky plane geometry they are curves of *geodesic curvature* (defined below) equal to 1 at every point, as we shall see soon.

If we transport these horizontal lines from $z = \infty$ to any point b of the absolute (by an isometry of the preceding theorem), then we get, in the Euclidean geometry terminology, the set of the circles tangent to the absolute at the common point b (Fig. 8.5).

These Euclidean circles tangent to the absolute, are called the *oricycles* of the Poincaré model of the Lobachevsky plane (of course, the horizontal lines $a = 0$, from which we started, are also oricycles in the Poincaré model).

Thus we have constructed a 1-parameter family of oricycles orthogonal to the "parallel" straight lines emanating from b.

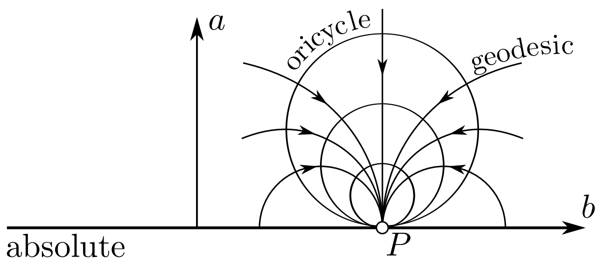


Figure 8.5: A family of parallel geodesics (straight lines) of the Lobachevsky plane, leading to the same point b of the absolute; and their orthogonal oricycles.

Problem. Calculate the length of a segment of oricycle between two fixed parallel geodesics, as a function of the position t of the oricycle point on one of these geodesics (where t is the distance along this geodesic from some fixed "initial point" of it).

Solution. We choose the family of parallel geodesics $b = \text{const}$ (between $b = 0$ and $b = c$). For the initial point e, $z = i$, on the geodesic $b = 0$, the distance to the point $z = Ai$ on that geodesic is equal to

$$t = \int_1^A \frac{da}{a} = \ln A,$$

that is, $A = e^t$.

Hence, the length of the segment of the oricycle passing through the point $z = Ai$, is equal to

$$\int_0^c \frac{db}{A} = ce^{-t} \ .$$

Corollary. *Two parallel geodesics approach each other exponentially, when one follows one of them with a constant velocity.*

Problem. Find the area of the "triangle" bounded by three straight lines of the Lobachevsky plane, that connect three points of the absolute, in the Poincaré model.

Solution. All such triangles are isometric —having then equal areas—, since one can send one to the other by an $\mathrm{SL}(2, \mathbb{R})$-isometry of the preceding theorem. Thus it suffices to study the triangle whose vertices are $z = -1$, $z = 1$, $z = \infty$, in the Poincaré model notations.

Answer. The area of this infinite "triangle" is finite, namely it is equal to π. The angles of this triangle are equal to 0.

The difference between π and the sum of the angles of a triangle is an additive function (since we add the angles with sum π at the dividing vertex).

As we shall see later, for any triangle of the Lobachevsky plane, the sum of the angles equals π *minus the area of the triangle.*

Thus, when the area is small, we approach the case of Euclidean geometry for which the sum equals π. The infinite triangle of the preceding problem is the opposite case: there exist no triangles of larger area (or of smaller angles).

Thus, the (non-oriented) area of any triangle in the Lobachevsky plane is a number between 0 and π.

Problem. Find the *equidistant curves* of the geodesic line $b = 0$ of the Poincaré model of the Lobachevsky plane.

Solution. The homotheties $z \mapsto tz$ $(t > 0)$ send every Euclidean straight line ray, containing 0, to itself, and are isometries of the Lobachevsky plane, preserving the geodesic line $b = 0$, but moving it along itself. Therefore, they move along itself every equidistant curve to this geodesic.

Hence, all the equidistant curves of this geodesic are represented in the Poincaré model by the Euclidean straight lines that meet this geodesic at the absolute. In fact, such lines meet also our geodesic at the point $z = \infty$ of the absolute.

Corollary. *The equidistant curves of the geodesic lines of the Lobachevsky plane are represented in the Poincaré model by the Euclidean circular arcs, joining the* 2 *intersection points of the original geodesic with the absolute* – Fig. 8.6.

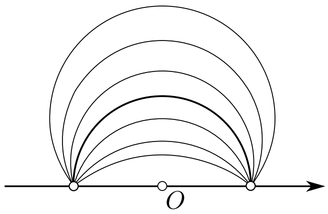


Figure 8.6: The equidistant curves of the geodesic line $a^2 + b^2 = 1$ (in the Poincaré model of the Lobachevsky plane).

Proof. Any geodesic line can be sent onto the special line $b = 0$ of the preceding problem by an $\mathrm{SL}(2, \mathbb{Z})$-isometry (of the preceding theorem), and the connecting Euclidean straight lines of the model are sent by this isometry to the connecting circles (since the shifts, the homotheties and the inversions are sending the family of all straight lines and circles onto itself). □

Problem. Find the isometries of the Lobachevsky plane, that preserve the point e (where $z = i$ in the Poincaré model).

Solution. The equation $(\alpha i + \beta)/(\gamma i + \delta) = i$ provides the answer $\alpha i + \beta = \delta i - \gamma$, that is, $(\alpha = \delta, \beta = -\gamma)$. Taking the condition $\alpha\delta - \beta\gamma = 1$ into account, we set the orthogonal matrix of a rotation at some angle ϑ

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{pmatrix} .$$

We conclude that there is a one-parameter family $\mathbb{S}^1 = \mathrm{SO}(2)$ of rotations of the Lobachevsky plane around e. The orbits of these rotations for the different angles ϑ, are closed curves, namely, they are the Lobachevsky circles, with "centre" e in the Poicaré model.

In fact these Lobachevsky circles (consisting of the points at a fixed distance from the central point) are represented in the Poincaré model by the Euclidean circles that do not intersect the absolute line.

However, their Euclidean centres are generically different from the centres in the Lobachevsky sense. Say, the Lobachevsky circle of radius R, with centre e $(z = i)$ intersects the diameter $b = 0$ at the points $z = e^{\pm t}i$, and hence the Euclidean centre of that circle is $z = i\ \mathrm{ch(t)} \neq \mathrm{i}$.

To see that *the Lobachevsky circles in the Poincaré model are represented by the Euclidean circles*, it suffices to send the upper half-plane $\mathrm{Im\ z} > 0$ to the disc $|w| < 1$ by a complex projective transformation, sending the point e $(z = i)$ to the centre $(w = 0)$.

The rotations of the Lobachevsky plane, described above, are represented in the coordinate w as the complex projective transformations that preserve the disc $|w| \leqslant 1$, and our rotations are represented by the ordinary rotations, $w \mapsto cw$, $c = e^{i\vartheta}$. Their orbits are the ordinary circles $|w| = r$ (Fig. 8.7).

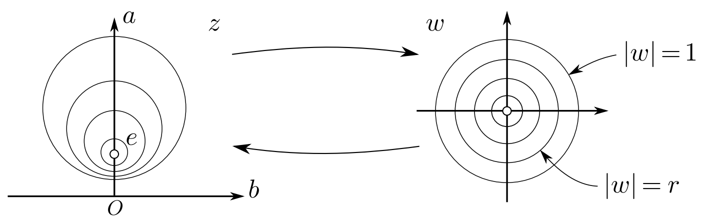


Figure 8.7: The family of concentric Lobachevsky circles centred at e (where $z = i$).

To return to the coordinate z, we have to make the inverse complex projective transformation, which sends these circles in the w-plane to some circles in the z-plane. But this transformation is a combination of shifts, of homotheties and of inversions, all of which respect the circles, as we have shown some pages above.

8.4 Parallel transport on Lobachevsky plane

As we shall see now, the circles, as well as the oricycles and the equidistant curves of the geodesic lines, are the only curves of the Lobachevsky plane that have constant geodesic curvature (defined below).

Namely, the geodesic curvature of a circle is greater than 1 (behaving asymptotically like $1/r$, when the Lobachevsky radius r tends to 0, and tending to 1, when the radius r tends to ∞).

The geodesic curvature of the equidistant curves is smaller than 1 (tending to 0 together with the distance from the geodesic line and to 1 when the distance grows to infinity).

A nice exercise is to find the explicit formulas for the geodesic curvature in terms of the Lobachevsky radius r and of the distance of the equidistant curve from the geodesic line.

We have to recall first the definition of geodesic curvature.

This definition depends on an extremely important (and rather nontrivial) notion of *parallel transport* of tangent vectors along Riemannian manifolds. We shall discuss this general notion in Chapter 9, defining it here only for the simpler case of the 2-dimensional manifolds, in order to apply it to the Lobachevsky plane – Fig. 8.8.

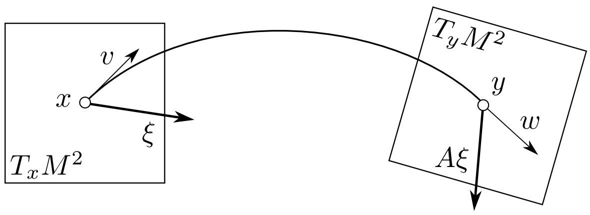


Figure 8.8: The parallel transport A of a tangent vector ξ along the geodesic line xy.

Definition. The *parallel transport* of a tangent vector $\xi \in T_xM^2$ along a geodesic line xy of the Riemannian surface M^2 is a linear map of the tangent space at the initial point to the tangent space at the final point

$$A : T_xM^2 \longrightarrow T_yM^2 \ ,$$

that preserves the lengths of the vectors, depends continuously on y, is the identity map for $y = x$ and sends the orienting tangent vector v of the geodesic xy at x to the tangent vector w of the same geodesic at the point y.

Example. The parallel transport of a vector tangent to the sphere of the Earth at Paris —directed to the East, and tangent to the parallel through Paris— to Barcelona (or to any place x at the same meridian as Paris) is a vector directed also to the East, but along the parallel through Barcelona (respectively along the parallel through x).

Now, to transport a vector $\xi = \xi_0$ along an arbitrary smooth curve γ connecting x to y (Fig. 8.9), we subdivide γ into small segments $\gamma_1, \dots, \gamma_N$ by the points $(x_0 = x, x_1, \dots, x_{N-1}, x_N = y)$ and then transport $\xi = \xi_0 \in T_{x_0}M$ to $T_{x_1}M$ along the short geodesic line connecting x_0 to x_1, then transport the resulting vector ξ_1 from $T_{x_1}M$ to $T_{x_2}M$ along the second short geodesic line connecting x_1 to x_2, and so on. After N steps we obtain a vector ξ_N at the final point $x_N = y$, in the space T_yM.

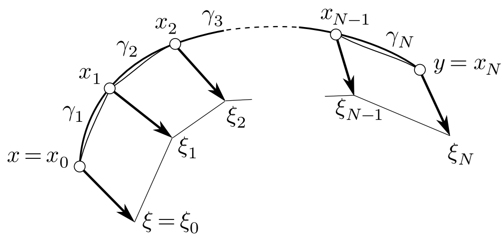


Figure 8.9: Parallel transport from x to y along a smooth curve as the limit of the parallel transports along the broken geodesic lines $(x_0, x_1, \dots, x_N)$.

This transport depends on N and on the subdivision, but for $N \to \infty$ and $|\gamma_j| \to 0$ there is a limit, which is called the parallel transport $A\xi$ of the vector ξ along the smooth curve γ. It defines a linear operator

$$A : T_xM^2 \longrightarrow T_yM^2$$

that preserves the lengths of the vectors, depends on y continuously, is the identity for $y = x$, but which, generically, sends the tangent vector to γ at x, to a non-tangent vector at the endpoint y.

Example. For the Euclidean plane, this parallel transport is really parallel: The components of the vectors ξ and $A\xi$ in a fixed linear coordinate system in $\mathbb{R}^2$ are equal. For other surfaces it is less evident.

Problem. Transport a tangent vector, directed toward the North pole at Saint-Petersburg (at latitude 60°) along its parallel moving to the East and returning back to Saint-Petersburg from the Occident. Calculate the direction of the transported vector.

This exercise helps a lot to understand the rather non-trivial preceding definition. However, a clever trick makes the calculation easier (in this particular example).

The initial (original) definition of the parallel transport was invented by the geodesists, and is well defined for the submanifolds of Euclidean space, whose Riemannian metrics are inherited from the embedding into this Euclidean space (which is the case of the Earth spherical surface in $\mathbb{R}^3$).

Now, to transport a vector ξ from a point x to an "infinitesimally close" point y of the submanifold M of $\mathbb{R}^n$, the geodesists proposal is to realise T_xM and T_yM as two affine planes in Euclidean space $\mathbb{R}^n$..

They project orthogonally ξ from the affine space T_xM to the affine space T_yM, transporting then the projected vector to the origin of T_yM by the parallel transport of this affine space.

To transport along the smooth curve γ, we subdivide it into many small pieces, use the projection for each piece and the resulting "broken transports along the submanifold M" converge to the parallel transport along γ.

It is not so difficult to prove that this process leads to the same parallel transport of our definition.

Corollary. *The limit of the broken transports along the submanifold M^2 of Euclidean space $\mathbb{R}^n$ does not depend on the embedding of the Riemannian submanifold M^2 into $\mathbb{R}^n$: It only depends on the interior Riemannian geometry of the surface, being the same for different isometric embeddings.*

Example. The limit of the broken transport along the surface of a cylinder or of a cone in $\mathbb{R}^3$ is the parallel transport along the corresponding curve on the Euclidean plane covering the cylinder or the cone isometrically.

Problem. Make a parallel transport of the vector directed toward the vertex O of the cone that intersects the plane $z = 1$ along the circle

$$x^2 + y^2 = r^2,$$

in the Cartesian coordinates (x, y, z) of Euclidean space $\mathbb{R}^3$ – Fig. 8.10.

Solution. The length of the circle is $2\pi r$, and the distance of its points from O is $R = \sqrt{1 + r^2}$. Thus, the plane sector of radius R, covering the cone, has the angle $\varphi = 2\pi(r/R)$.

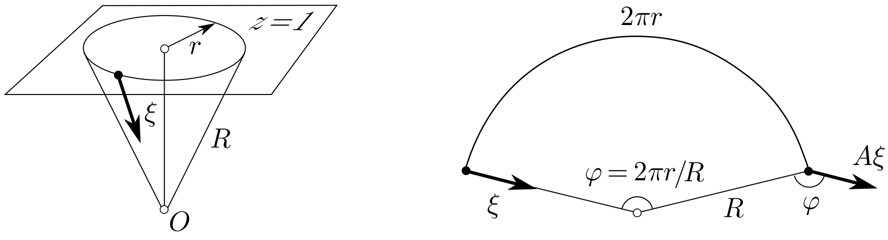


Figure 8.10: Parallel transport on a cone.

The parallel transport along any curve on the conical surface is the transport in Euclidean plane along the image of this curve on our sector.

Hence, the transported vector $A\xi$ will form the angle φ with the initial vector, as it is shown in Fig. 8.10. Consequently, for $r = \sqrt{3}$ we get $R = 2$, $r/R = \sqrt{3}/2$, $\varphi = \pi\sqrt{3}$ and $2\pi - \varphi = \pi(2 - \sqrt{3}) \approx 0,27 \cdot \pi \approx 50°$.

This result is directly applicable to the Saint-Petersburg problem. The cone one has to use is formed by the tangent straight lines to the meridians at all the points of the parallel of latitude 60° (Fig. 8.11).

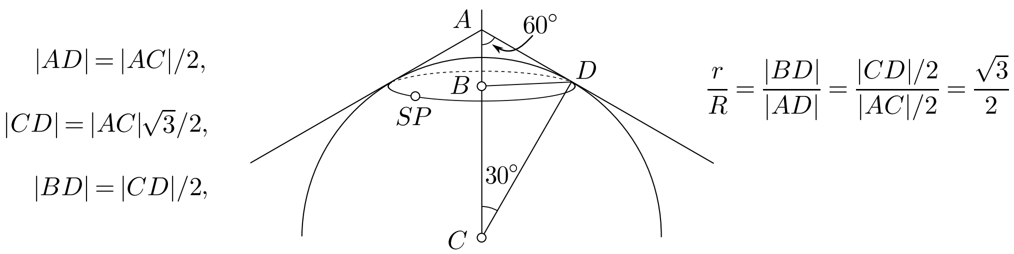


Figure 8.11: The cone tangent to the Earth along the Saint-Petersburg parallel.

The parallel transports along this parallel, obtained from the broken transport construction, coincide for the sphere and for the cone surface, since both surfaces have the same tangent affine planes at the points of the parallel.

Hence it suffices to calculate the ratio r/R for the latitude 60°. It is easy to see that it is $\sqrt{3}/2$. As we have seen, the tangent vectors turns by 50°.

Problem. Find the parallel transport of the vertical vector ($\dot{a} = 1, \dot{b} = 0$) along the oricycle $a = 1$ of the Poincaré model of the Lobachevsky plane, from the point $z = i$ to the point $z = i + t$ at distance t.

Solution. Subdivide the oricycle into N small segments of length t/N, and make the parallel transport along the geodesics connecting two neighbouring division points x_k and x_{k+1}. We observe (Fig. 8.12) that this geodesic line is represented in the Poincaré model by a Euclidean circle of Euclidean radius $\varrho = 1 + O(1/N^2)$, and that its directions at the two division points are inclined (in the Euclidean geometry) at an angle $\varphi_k = 1/N + O(1/N^2)$.

Therefore, the broken transport along the small geodesic lines provides the angle of rotation $t + O(1/N)$ in the Euclidean geometry, and the limiting rotation angle is thus t, as shown in Fig. 8.12.

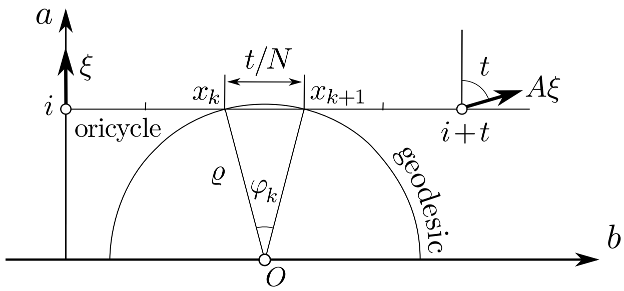


Figure 8.12: The parallel transport of the vector $\xi \in T_iG$ to the vector $A\xi \in T_{i+t}G$ along the oricycle $a = 1$ of the Lobachevsky plane G.

8.5 Geodesic and Gaussian curvatures

Definition. The *geodesic curvature* of a smooth curve on a Riemannian surface is the speed of change of the angle φ between a vector transported parallelly along this curve and the tangent vector to the curve, moving along the curve with speed 1, $k = d\varphi/ds$ (φ being defined up to a constant).

It measures the local speed of deviation of the curve from its tangent geodesic.

Example. The curvature of a curve in Euclidean plane is $k = d\varphi/ds$, where s is the length parameter along the curve and φ the angle of the tangent vector with a fixed vector of the plane (since parallel transport is usual parallelism).

PROBLEM. Find the geodesic curvatures of the oricycles of Lobachevsky plane.

SOLUTION. All the oricycles have equal geodesic curvatures, constant at all their points. This is because there are isometries (and we know them) that send each oricycle to any other and isometries that send any point of the oricycle $a = 1$ to any other point of it.

For this special oricycle, the solution of the preceding problem shows that moving at a distance t along the oricycle from the original point, the angle between the parallelly transported vector $A\xi$ and the tangent vector to the oricycle at the point $i+t$, is $\varphi = t$. In consequence,

$$k = \frac{d\varphi}{dt} = \frac{dt}{dt} = 1,$$

and hence all the oricycles have everywhere geodesic curvature 1.

Problem. Find the geodesic curvature of a Lobachevsky circle of radius r.

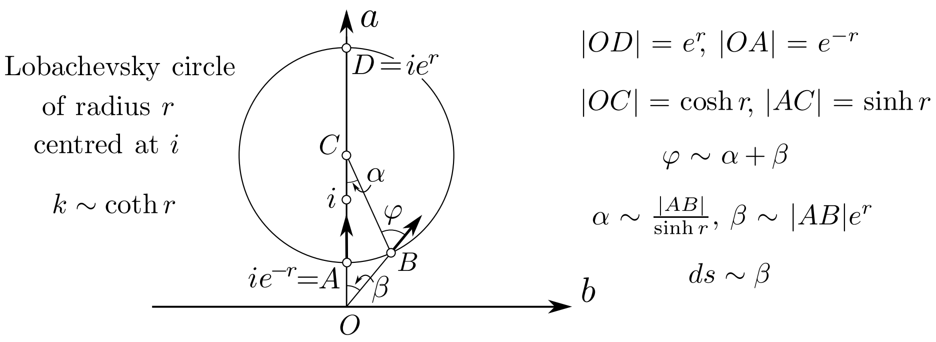


Figure 8.13: The infinitesimal parallel transport of vertical vector v from A to B along the Lobachevsky circle of radius r, centred at $z = i$.

Solution. All these circles have constant geodesic curvature everywhere, since they are isometric (to each other) and can be rotated by some isometries. Thus, we shall consider the Euclidean circle intersecting the axis $b = 0$ at the points $D : z = ie^r$ and $A : z = ie^{-r}$, whose Euclidean centre C is $z = i\ \mathrm{ch}(r)$ – Fig. 8.13.

To compute the geodesic curvature of this Lobachevsky circle of radius r at A, we transport the vertical vector $v \in T_A G$ parallelly to the (infinitesimally) close point B of the circle. The corresponding geodesic line being (in first approximation) the Euclidean circle centred at O, we find the transported vector, continuing the direction OB. The Lobachevsky normal vector to the circle has the direction of the Euclidean vector BC, since it is orthogonal to the circle also in the Euclidean sense. Therefore, the rotation angle is $\varphi = \alpha + \beta$ (in the first differentials approximation), the Lobachevsky distance from A to B being the Euclidean distance divided by $|OA|$, that is β.

Finally, $\alpha/\beta \sim |OA|/|CA|$ (for the Euclidean distances), $|OA| = e^{-r}$, $|CA| = \mathrm{ch}(r) - e^{-r} = \mathrm{sh}(r)$. Therefore the geodesic curvature of our circular curve at A is

$$\frac{d\varphi}{dt} \sim \frac{\alpha+\beta}{\beta} = 1 + \frac{e^{-r}}{\mathrm{sh}(r)} = \frac{\mathrm{sh}(r)+e^{-r}}{\mathrm{sh}(r)} = \frac{\mathrm{ch}(r)}{\mathrm{sh}(r)} = \mathrm{cth}(r)\ .$$

For $r \to 0$ it behaves like $1/r$, as the curvature of the Euclidean circle of radius r. For $r \to \infty$ the geodesic curvature $\mathrm{cth}(r)$ of a Lobachevsky circle of radius r tends to 1, which is the geodesic curvature of an oricycle – an oricycle can be considered as a Lobachevsky circle of radius $r = \infty$.

Gaussian Curvature. The notion of geodesic curvature leads to the important differential 2-form on a Riemannian surface, called "curvature form".

First, given a curve $\gamma : [s_0, s_1] \to M$ on a surface, we take the product of the geodesic curvature with the length element, as a differential form $k\,ds$ $(= d\varphi)$ along the curve. Then the integral of the geodesic curvature along the curve is the angle between the transport along γ of the tangent vector to γ at $\gamma(s_0)$ (to $\gamma(s_1)$) and the tangent vector to γ at $\gamma(s_1)$: $\Delta\varphi = \int_\gamma k\,ds$.

Observe that the sign of the geodesic curvature depends on the orientation of the surface, which determines the sign of the angle φ involved in the definition.

Next, we consider a small oriented piece σ^2 of the surface M and let $\gamma = \partial\sigma^2$ be its oriented boundary – in this case the curvature sign is defined by the convention of the orientations of σ^2 and of γ.

Example. For the circle of radius r in Euclidean plane the (positive) geodesic curvature is $k = 1/r$, while for the circle of radius r in Lobachevsky plane the geodesic curvature is

$$k = \mathrm{cth}(r) = \frac{e^r + e^{-r}}{e^r - e^{-r}} \approx \frac{1}{r}\,(1 + \frac{r^2}{3} + \ldots)\ .$$

The angle of rotation of the tangent vector minus the integral of the geodesic curvature along the boundary curve $\partial\sigma^2$, $2\pi - \int_{\partial\sigma^2} k\,ds$, is an additive function of the 2-chain σ^2. Hence it can be expressed as the integral of some differential 2-form ω along the chain. But since any 2-form is a multiple of the area element $d\sigma$, this 2-form is written as $\omega = K\,d\sigma$, where K is a function and $d\sigma$ is the Riemannian oriented area element form.

Definition. This function K is called the *Gaussian curvature* of the Riemannian surface M^2, and $\omega = K\,d\sigma$ is called its *curvature form.*

Example. The Gaussian curvature of a Euclidean plane (embedded or not in $\mathbb{R}^3$) is $K = 0$. Indeed, the boundary of any small disc Δ_r of radius r has geodesic curvature $k = 1/r$ and length $2\pi r$, which gives $\int_{\partial\Delta_r} k\,ds = 2\pi$. Hence

$$\int_{\Delta_r} K d\sigma = 2\pi - \int_{\partial\Delta_r} k\,ds = 0,$$

that is, $K = 0$.

Exercise. Verify that the Gaussian curvature of a sphere of radius R in Euclidean space $\mathbb{R}^3$ is $K = 1/R^2$.

For a surface M^2 embedded in Euclidean space $\mathbb{R}^3$, which inherits its Riemannian metric from that embedding, the Gaussian curvature K defined here coincides with the product of the principal curvatures – or with the Jacobian determinant of the Gauss map $M^2 \to \mathbb{S}^2$ defined by the embedding (see Ch. 9, p. 315-316, where these products and Jacobians are discussed). This coincidence is proved in Ch. 9, where we use another (equivalent) definition of the Gaussian curvature function K. The advantage of the present definition and of that of Ch. 9 is their independence of any embedding: Two isometric surfaces have equal Gaussian curvatures at the points sent one to the other by an isometric map of one surface to the other.

It follows, in particular, that the "exterior" Gaussian curvature of an isometrically embedded surface is independent of the embedding. Moreover, the general Gauss-Bonnet formula

$$\iint\limits_{M^2} K\,d\sigma^2 = 4\pi(1-g)$$

holds for any surface M^2 of genus g. This identity holds independently of any embedding (be it isometric or not). We shall prove it in Ch. 10, p. 376-377, for the standard embeddings of M^2 in Euclidean space $\mathbb{R}^3$.

Problem. Calculate the Gaussian curvature of the Lobachevsky plane.

Solution. We take a small square σ^2, bounded by the lines $a = 1$ and $a = 1 + \varepsilon$, $b = 0$ and $b = \varepsilon$ in the Poincaré model. Its boundary is not a smooth curve, and at each angular point the direction φ of the curve turns $\pi/2$. But to compute the integral of $k\,ds$, we may use this square with the following correction: We shall add, to the integrals of $d\varphi$ along the sides, the jump of φ $+\pi/2$ at each angle.

The vertical sides (b = const) are geodesic lines, and hence there is no geodesic curvature. The geodesic curvatures of the horizontal sides (a = const) are equal to 1, since they are oricycles. The rotations along the two horizontal sides make to turn the tangent vectors in opposite directions, and the rotation angles are, up to their signs, the lengths of the sides, that is $-\varepsilon$ for $a = 1$ and $-\varepsilon/(1+\varepsilon)$ for $a = 1+\varepsilon$. The total angle of rotation $\int d\varphi$ along the horizontal sides is the difference $-\varepsilon + \varepsilon/(1+\varepsilon) = -\varepsilon^2 + o(\varepsilon^2)$.

The total contribution of the 4 angles, $4(\pi/2) = 2\pi$, does not turn the vectors, and hence

$$\iint_{\sigma^2} K\, d\sigma^2 = -\varepsilon^2 + o(\varepsilon^2)\ .$$

The Lobachevsky area of our small chain is $-\varepsilon^2 + o(\varepsilon^2)$. In consequence, the Gaussian curvature K of the Lobachevsky plane is equal to -1.

Lobachevsky plane as "complex sphere". The Gaussian curvature of the sphere $\mathbb{S}^2$ of radius R is equal to R^{-2}. Thus one proclaims the Lobachevsky plane to be "the sphere of imaginary radius $R = i$".

Problem. Prove that the three medians of a triangle in the Lobachevsky plane have a common intersection point.

Solution. For the Euclidean plane triangles it is well-known (it is an easy consequence of the vector algebra).

For the spherical triangles on $\mathbb{S}^2 \subset \mathbb{R}^3$, it is also the case. It suffices to project the spherical triangle ABC to the plane ABC by the rays from the origin to send the spherical triangle medians to the plane triangle medians.

Considering the Lobachevsky plane as a sphere of imaginary radius, and writing the geometrical theorems as algebraic identities, we continue these identities to the complex values of the arguments (since an algebraic identity, like $(a+b)^2 = a^2+2ab+b^2$, remains true for the complex variables, whenever it is true for all real values of the variables). Since the medians intersection property is true for the spheres of real radius, it should persist on the complex spheres, including the Lobachevsky plane.

Of course, we do not insist on the mathematical rigour of this proof (suggested by the physicists). However, the theorem is true, and the missing points of the proof may be correctly performed with all the mathematical rigour (but we leave to the readers the pleasure to add these missing details).

Problem. Have the altitudes of any Lobachevsky plane triangle a common intersection point?

We shall see that the complex version of this intersection property is true, but the real intersection points of the real Lobachevsky plane live sometimes outside the Lobachevsky plane: They live in a relativistic version of the Lobachevsky plane, which is different from it in the real form, but coincides with it in the complexified version. See Fig. 8.14, where the three altitudes have no intersection – this happens only when one of the angles of the triangle is greater than 120°.

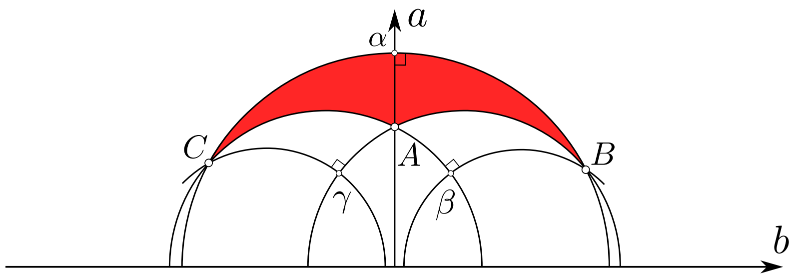


Figure 8.14: The three altitudes $(A\alpha)$, $(B\beta)$ and $(C\gamma)$ of the Lobachevsky triangle ABC do not intersect each other.

Problem. Prove the theorem on the sum of the angles of a triangle in the Lobachevsky plane:

$$\alpha + \beta + \gamma = \pi - S \ , \tag{2}$$

where S is the area of the triangle.

Hint. Integrate the form "$d\varphi$"$= k\,ds$ along the triangle. The total integral is $KS = -S$, since the Gaussian curvature of the Lobachevsky plane is $K = -1$ everywhere.

The whole contribution to the integral is provided by the three vertices, since the geodesic curvature k of the sides is equal to 0. The contribution of the interior angle α to the rotation of the tangent vector is $\pi - \alpha$, and we get finally

$$(\pi - \alpha) + (\pi - \beta) + (\pi - \gamma) = -S \pmod{2\pi} \ ,$$

whence the formula (2) follows. The integral multiple of 2π to add to the term $-S$ should be independent of the triangles (by continuity), and it is 2π for the small triangles, by the Euclidean geometry theorem: $\alpha + \beta + \gamma \to \pi$ for $S \to 0$.

8.6 Poincaré model as optical medium

Our Euclidean knowledge that the shortest path between two points is a segment of straight line and Galileo Inertia Law, stating that a moving body will continue in straight-line motion (unless acted upon by an external force), both are intimately associated to our visual perception of the trajectory of light rays, that is, of (what we consider) a straight line. This is because in a *homogeneous optical medium* (in which the light velocity is the same at all its points) light propagates in straight lines.

In an non-homogeneous medium, the velocity of light depends on the points of this medium, $v = v(x)$, and light rays become curvilinear. Hence, in general, objects are not placed in the direction perceived by the observer. These facts are essential for astronomers, for sailors and for the constructors of all kind of optical devices, including the instrumentation of guided missiles. In order to understand curvilinear rays and to provide a natural Riemannian metric to an optical medium, let us recall the law of refraction.

Snell law. The deviation of light from a straight line as it crosses the interface C between two homogeneous optical media M_1, M_2 (for example air-water) is called *refraction.* Snell law states that the ratio of the respective light velocities v_1, v_2 (in the two media) is equal to the ratio of the sines of the angles of incidence and refraction, $\sin\alpha_1/\sin\alpha_2 = v_1/v_2$ (Fig. 8.15). Moreover, the incidence and refracted rays are coplanar with the normal to the interface.

$$\frac{\sin\alpha_1}{v_1} = \frac{\sin\alpha_2}{v_2}$$

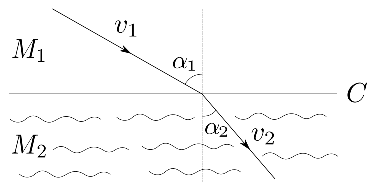


Figure 8.15: Snell law of refraction.

Main Researches on the Refraction Law. For several centuries astronomers and mathematicians tried to discover the law of refraction. In the 2nd century, Ptolemy, neglecting his data to fit theory, had found an inaccurate "empirical law" on the refraction angles. In 984, the law of refraction was described by Ibn Sahl (of Baghdad), who used it to work out the shapes of lenses. Later, in 1602, Thomas Harriot (of Oxford) rediscovered the law without publishing it (although he mentioned the law in his correspondence with Kepler). After the death of the Dutch astronomer and mathematician Willebrord Snellius (Snell), in 1626, among his papers was found a work of 1621 in which the law of refraction was stated from experiments. In 1637 the French mathematician and philosopher René Descartes published, in his book *Dioptrics*, a "theoretical deduction" of the law of refraction based on two erroneous assumptions: 1. That velocity increases when light passes from air to a more dense optical medium; and 2. When light passes from a medium to another, only the component of the velocity normal to the interface varies. In accordance with his conception of science as a hypothetical-deductive system, Descartes preferred sometimes to follow his general principles rather than to comply with the experimental results. In 1678, the Dutch physicist and mathematician Christiaan Huygens, who considered light as a wave propagation process, deduced Snell law from his method to describe wave propagation, known now as *Huygens principle* (a contact geometry description of Huygens principle is given in p. 618). In 1662, before Huygens, the French mathematician Pierre de Fermat proposed a general principle allowing to explain the light rays trajectory in different situations, included when light crosses the separation interface between two media. The *extremal principle of Fermat* states that: *the path followed by a light ray between two points is the path that can be traversed in the least time among all possible paths joining those points.* Fermat deduced the Snell law from his principle. Now it is known that Huygens principle implies Fermat principle.

Exercise. Deduce the Snell law from Fermat principle. Hint: Given two points A and B in the respective media M_1 and M_2, compute the time T_{AOB} needed to follow a broken line AOB as a function of a coordinate x on the interface (Fig. 8.16). Find its minimum.

$$T_{AOB} = \frac{\sqrt{a^2+x^2}}{v_1} + \frac{\sqrt{b^2+(h-x)^2}}{v_2}$$

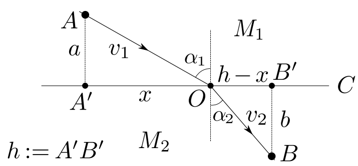


Figure 8.16: Deduction of the Snell law from Fermat Principle.

Parallel homogeneous strips. Consider a plane domain separated by horizontal parallel lines into optically homogeneous strips M_k, with light velocity v_k. Given two points A in M_1 and B in M_N, a light ray going from A to B is a broken line with vertices in the separating lines (Fig. 8.17 left).

Denote by ϑ_k the angle between the parallel lines that bound the strip M_k and the segment of ray that passes through this strip (ϑ_k is complementary to the incidence and refraction angles α_k in M_k, $\sin\alpha_k = \cos\vartheta_k$). Applying the Snell law to each consecutive interface we get a sequence of equalities whose common constant value we denote by $\varkappa$

$$\frac{\cos\vartheta_1}{v_1} = \frac{\cos\vartheta_2}{v_2} = \ldots = \frac{\cos\vartheta_N}{v_N} = \varkappa = \text{const.}$$

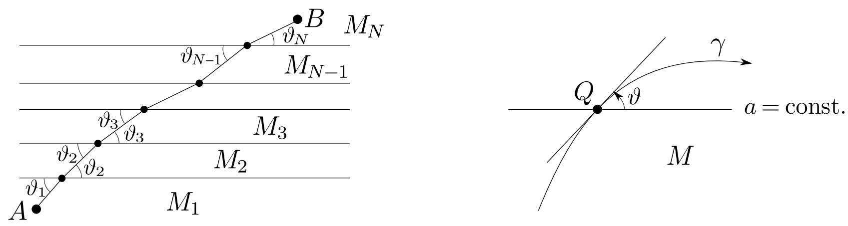


Figure 8.17: Left: Homogeneous strips; Right: Along a light ray we have $\frac{\cos\vartheta}{v} = \varkappa$.

Therefore along the light ray, except at its vertices, the relation

$$\frac{\cos\vartheta}{v} = \varkappa$$

holds, where ϑ is the angle formed by any light ray segment with the horizontal direction and v is the velocity of light along that segment.

Curved light rays. Consider a plane optical medium M that is homogeneous along the horizontal layers $a = \text{const}$. That is, we assume the velocity of light is a continuous function of the ordinate $v = v(a)$.

Proposition. *In such a medium, along a light ray γ the preservation law*

$$\frac{\cos\vartheta}{v} = \varkappa = \text{const.} \tag{3}$$

holds, where v is the velocity of light at any point Q of γ and ϑ is the angle formed by the tangent to γ at Q with the horizontal layer (Fig. 8.17 right).

To prove (3) we shall modify slightly the distribution of velocities of light in M to get a medium M_ε formed by horizontal homogeneous strips of width less than ε, conveying that the velocity of light inside each strip is equal to the velocity of light in an interior layer of the given strip. As we have seen, a light ray in this medium M_ε is a broken line along which equation (3) holds.

In the limit, when the width ε tends to zero, we get the initial continuous distribution of the velocities of light in M, and each broken light ray in M_ε tends to a light ray in M for which relation (3) holds (Fig. 8.17 right).

Corollary. *In any such medium, all vertical lines are light ray paths.*

Surfaces of Revolution. Consider a surface of revolution around the z-axis in Euclidean space $\{(x,y,z)\}$. Its geodesics satisfy an equation similar to (3) in which $1/v$ is replaced by distance $r(z)$ to the rotation axis:

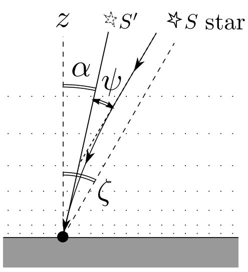
Clairaut's Theorem. *The product of the distance to the axis of rotation and the cosine of the angle a geodesic makes with a parallel is constant along each geodesic of a surface of revolution*

$$r \cos \vartheta = \text{const.} \tag{4}$$

Corollary. *In a surface of revolution all meridians are geodeics.*

Atmospheric Refraction. Assuming a flat earth, the atmosphere is an optical medium for which, at first approximation, the velocity of light depends on the height, $v = v(y)$, and hence relation (3) holds along light rays. Indeed, knowing that atmospheric pressure decreases with height (proved by Pascal in 1648) and that "air density is proportional to the weight of all covering air" Newton concluded that air density decreases with height according to an exponential law $\rho = \rho_0 e^{\beta y}$, (while velocity of light increases).

If there was no refraction, a celestial body would be seen under an angle ζ with the vertical, called *zenith angle*. But due to the atmospheric refraction the body is seen under an *apparent zenith angle* $\alpha < \zeta$, so that the body seems to be closer to the zenith than it really is (Fig. 8.18 *a*). The difference



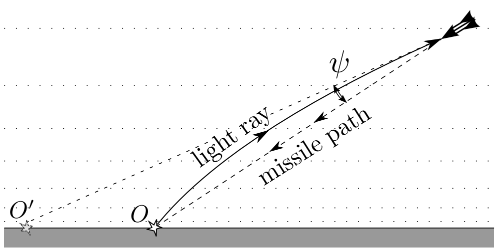


Figure 8.18: *a*) astronomical refraction $\psi = \zeta - \alpha$; *b*) terrestrial refraction ψ.

$\psi = \zeta - \alpha$ is called *astronomical refraction*; it is zero at the zenith and increases with increasing apparent angle α, attaining $35'$ at the horizon ($\alpha = 90°$). So, in a sunset, when the lower edge of the sun "touches the horizon", this edge is in reality $35'$ below the horizon; the whole solar disk is also below because it determines an angle of $32'$ and, moreover, its shape is perceived flattened in the vertical direction because refraction raises the upper edge a few minutes of arc less than the lower one.

Terrestrial refraction is the deviation of rays travelling from objects located in the atmosphere. Newton noticed that several phenomena (humidity, winds, temperature changes, etc.) produce air density variations which modify the refractivity, specially in low atmospheric layers. Modern techniques to measure refractivity are unfortunately used for optically guided missiles (Fig. 8.18 b).

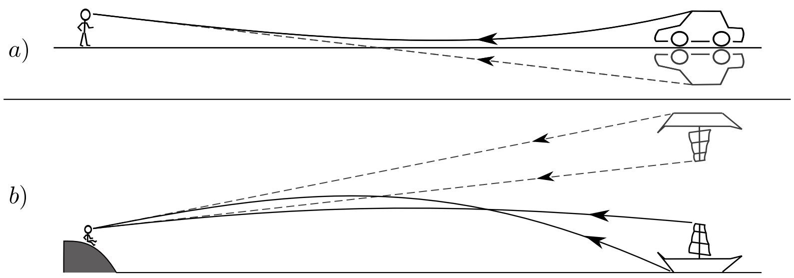


Figure 8.19: *a*) hot-road inferior mirage; *b*) cold-sea superior mirage.

The "mirages" (occurrence of erect or inverted images of the same object) are explained by the fact that *the convexity of a curvilinear light ray is directed to the part where density is decreasing* (Fig. 8.19).

The optical length of a curve in an optical medium is the time which is needed to follow this curve at the speed of light. Optical length induces a metric in the medium and, according to the extremal principle of Fermat, the light rays are the paths of extremal optical length (the geodesics). Conversely, we can think a Riemannian manifold as an optical medium whose light rays are the geodesics. In this way, the atmosphere can be considered as a Riemannian domain whose (optical) Riemannian metric is permanently changing.

Optics in Lobachevsky plane. Consider the half plane as an optical medium in which the velocity of light is equal to the ordinate, $v(a, b) = a$.

Theorem 5. *The light rays in Poincaré half plane are its geodesics.*

Proof. A curve is a light ray if and only if relation (3) holds on it with $v = a$:

$$\frac{\cos\vartheta}{a} = \varkappa. \tag{5}$$

We shall prove that a curve satisfies (5) if and only if it is a geodesic (a circle centred at the absolute ($a = 0$) or a vertical line). Let γ be a circle of radius R centred at a point O on the absolute and Q any point of γ (Fig. 8.20).

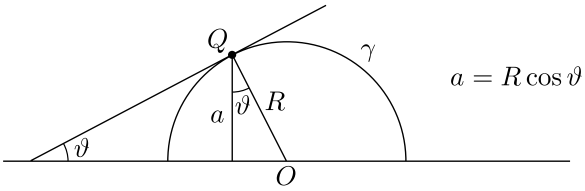


Figure 8.20: If velocity of light is $v(a,b) = a$, light rays are Lobachevsky geodesics.

Write a for the ordinate of Q and ϑ for the angle formed by the tangent to γ at Q with the horizontal direction. Then $a = \pm R\cos\vartheta$, that is

$$\frac{\cos\vartheta}{a} = \varkappa\,, \qquad \text{where} \quad \varkappa = \pm\frac{1}{R}$$

and the sign $\pm$ depends on the orientation of γ (giving the sign of $\cos\vartheta$). □

So the Euclidean curvature of the light ray, $\varkappa = \pm 1/R$, is the momentum of a free point, which is the constant of motion $p_b = \varkappa = \pm 1/R$ mentioned in the remark of p. 273. Indeed $p_b = \dot{b}/a^2$ and, since $v = a$, the horizontal component of the velocity is $\dot{b} = a\cos\vartheta$ implying that $\varkappa = p_b = \cos\vartheta/a$.

Newton, Flamsteed and Halley. Newton's researches on atmospheric refraction were part of his correspondence with Flamsteed (the best practical astronomer at that time), which turned into a quarrel. Here is a very small abstract of this story.

John Flamsteed (1646-1719) wrote his first paper on astronomy at the age of 19, beginning systematic observations in 1671. He was appointed the first English Astronomer Royal in 1675 and the Royal Observatory at Greenwich was built for his observations, where he spent more than forty years observing and making meticulous records. In 1676 he was admitted a Fellow of the Royal Society.

Edmond Halley (1656-1742). At the age of 19 he worked with Flamsteed, who showed his praise for him in a publication. At the age of 21 Halley was Captain of the first British ship built for scientific expeditions, publishing then the first "Catalogue of the Southern Stars". He was elected a member of the Royal Society at the age of 22.

The fame acquired by Halley turned Flamsteed's praise for Halley into jealousy, which soon became hatred. Their divergent personalities enhanced this enmity: In contrast to the strict christianity and secluded life of Flamsteed, who devoted all his time to astronomical observations, Halley doubted the biblical story of the creation and was know for his sense of humour and his great charm of manner. Halley travelled a lot and was frequently presenting new theories and discoveries in the Transactions of the Royal Society.

Having a close friendship with Newton, Halley stimulated and convinced him to write the "Principia Mathematica", and was very involved to carry out its printing. He paid for all the costs of publication out of his own pocket and he corrected the proofs. Soon after (1691) Flamsteed succeeded in preventing Halley's appointment as professor in Oxford.

Preparing the 2nd edition of his "Principia", Newton needed Flamsteed observations to perfect his Lunar theory and to have precise practical confirmations of his Gravitation Theory, which still was not accepted in continental Europe (where most scholars were explaining planetary motions by means of a peculiar *theory of vortices* of Descartes). Thus in September 1694, Newton visited Flamsteed who delivered him 50 observations.

They begun a correspondence in which Newton developed the theory of astronomical refraction and sent him (in 1695) his astronomical tables of refraction (published in 1721).

Flamsteed was however troublesome towards Newton since he felt that his observations did not receive sufficient credit in the previous lunar theory of Newton and due to Newton's closeness with Halley, who also needed his observations. Their relations were increasingly deteriorated due to Newton's arrogance (according to Flamsteed) and to Flamsteed's refusal to provide him his observations, saying that they were still not verified.

Flamsteed was mocking at Newton's ideas on gravity and was calling him "our great pretender". It seems that, having very limited mathematical knowledge and no interest on physical principles, Flamsteed has never understood Newton's concepts nor why Newton was needing more observations. But as his observations were urgently needed by Newton and Halley, in 1704, Prince George ordered to inspect Flamsteed's manuscripts and to proceed to print them, undertaking the publication expenses. The print process was going very slowly and, in 1708, was stopped by the Prince's death. Meanwhile, Newton managed in expelling Flamsteed from the Royal Society when his membership expired (1709).

In 1710, Newton got from Queen Anne an order to constitute a Council (presided by him) to demand to Flamsteed a copy of his observations. The Council proceeded to print them under Halley's direction and, despite Flamsteed's furious objections and refusal to provide some data, 400 copies of Flamsteed's Star Catalogue "Historia Coelestis" were finally printed in 1712, including a preface where Halley attributed the printing slowness to the sluggishness, secretiveness, and lack of public spirit of the author.

Queen Anne died in 1714 and Lord Halifax, Newton's supporter at court, in 1715. Then, having the support of Lord Chamberlain at court, Flamsteed managed to gather 300 copies that had not been distributed, burning them in Greenwich Park.

After Flamsteed's death, in 1719, Halley succeeded him as Astronomer Royal.

The story was re-opened in 1832 when, in the Royal Observatory, lots of letters belonging to Flamsteed were found. In some letters Newton expounded his atmospheric refraction theory, till then unpublished. English Admiralty published the letters in 1835.

Light and Matter. Light is a fascinating subject in which geometry has always played an important role. For example, contact geometry (discussed in Ch. 16) is the natural setting to describe wave propagation. Besides the above descriptions of light propagation, considered as rays that obey Fermat's principle or considered as a wave propagation process that obeys Huygens' principle (both very useful for lots of applications), we shall add few words on the subsequent studies of light.

In 1873 Maxwell unified electricity, magnetism and optics, into a theory of electromagnetic waves, all propagating at the velocity of light. In 1888, the experiments of Heinrich Rudolf Hertz confirmed Maxwell theory, which became accepted by all physicist.

However, some contradictions arose with experiments concerning momentary interaction between light and matter (in the photoelectric effect). To resolve those contradictions, Einstein, in his modestly named paper "On an heuristic point of view concerning the generation and transformation of light" (1905) proposed that *the energy of light is not distributed evenly over the whole wave front, as the classical theory assumed, but rather is concentrated on basic quantities localised in discrete small regions.* Such a small basic quantity of light-energy (a *quantum* of light) is now called photon (we can say *particle of light*).

It became later clear that all the electromagnetic waves are made of photons [Radio (Hertzian) waves, micro waves, infrared waves, visible light, ultraviolet light, X-rays, Gamma-rays, etc., whose only difference is their frequency]. Photon was not more a "heuristic point of view", becoming a fundamental part of science.

In 1929, a number physicist developed a theory on the interaction of light and matter (between photons and electrons), calling it "Quantum Electrodynamics" (QED) to frighten the non-specialists. It is often said that classical theory of light remains completely valid and most appropriate for dealing with the ordinary problems of optics, whereas QED must be used to understand the interaction between light and individual atoms. This is not completely true: QED predicts some striking optical phenomena that contradict the classical theory and "common sense" (for example, breaking reflection law), but that fully agree with experiments (see [69], a very nice and non-technical introduction to QED).

8.7 Klein model of Lobachevsky plane

We shall explain a completely different model of Lobachevsky geometry. It starts from the quadratic form $f(x,y,z) = z^2 - x^2 - y^2$, defining a Riemannian metric on the two-sheets hyperboloid M^2 given by the equation $f(x,y,z) = 1$ (Fig. 8.21). To define the metric, we have to construct a family of ellipses in

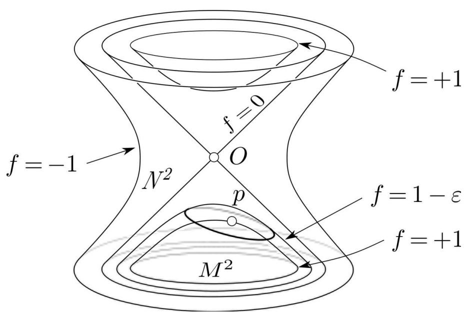


Figure 8.21: The Lobachevsky metric on the two-sheets hyperboloid surface, $f(x,y,z) = 1$, constructed from quadratic form $f = z^2 - x^2 - y^2$.

the tangent spaces to the surface M^2:

Consider the similar (and homothetic) deformed surface, $f = 1 - \varepsilon$. Its intersection with the affine tangent plane to M at p is an ellipse in this affine plane. In this way, we obtain a family of ellipses that define positive definite quadratic forms on the tangent spaces to M^2. To make them independent of ε, one should divide by $\sqrt{\varepsilon}$ the tangent vectors to M^2 that lie in the affine tangent plane and whose end-points belong to the ellipse, and then consider the limit tangent vectors for $\varepsilon \to 0$. The limit ellipses are independent of ε. These operations reduce the quadratic forms to the restrictions of minus the second differential of the function f to the tangent planes to M^2 (where $df = 0$).

Note that the resulting Riemannian metric on M^2 is invariant with respect to those linear transformations of the vector-space $\mathbb{R}^3$ which preserve the quadratic form f. Indeed, such a map acts on M as a diffeomorphism of M^2 onto itself, sending to itself also the perturbed surface $f = 1 - \varepsilon$. The tangent affine spaces of M^2 are sent to the tangent affine spaces, and their intersections with the perturbed hyperboloid $f = 1 - \varepsilon$ are sent to those intersections at the corresponding new points of M^2.

In consequence the Riemannian metric is sent to itself.

Definition. The hyperboloid M^2, equipped with this metric, is called the *hyperbolic model of Lobachevsky plane.*

As we shall see soon, this surface is isometric to the Poincaré model of the Lobachevsky plane.

Remark. Of course, the surface M is not connected: It has two connected components ($z > 0$ and $z < 0$). The two components are isometrically sent one to the other by the central symmetry sending (x, y, z) to $(-x, -y, -z)$.

To avoid this doubling, one may either suppose, say, $z > 0$ and forget the second component, or one may projectivise, considering the surface

$$\widetilde{M} = M^2/\{+1, -1\} \ .$$

Practically, one prefers to project the hyperboloid M^2 to the real projective plane $\mathbb{RP}^2 = (\mathbb{R}^3 \smallsetminus 0)/(\mathbb{R} \smallsetminus 0)$, which is the space of the straight lines containing the origin.

To represent the projective plane, one uses the affine coordinates on the plane $z = 1$. Thus, one represents the point (x, y, z) of M^2 by its projection $(X = \frac{x}{z}, Y = \frac{y}{z})$ into the affine plane $z = 1$ with coordinates X and Y. This projection cover the open unit disc $X^2 + Y^2 < 1$, so that each point of the disc represents one point of the upper sheet of M^2 and one (opposite) point on the lower sheet.

Therefore, the projection to this disc eliminates the unnecessary doubling of the hyperboloid model, and one usually prefers to use rather this disc model than the half-hyperboloid, diffeomorphic to it.

The resulting model of Lobachevsky plane in the disc $X^2 + Y^2 < 1$ is usually called "*Klein model*", since it was introduced by Arthur Cayley.

Theorem 6. *In the Klein model of Lobachevsky plane, the straight lines are represented by the Euclidean chords of the disc representing the plane.*

Proof. The linear map $(x, y, z) \mapsto (-x, y, z)$ preserves the form f and acts on M^2 as a reflection that preserves the hyperbola $x = 0$, $(z^2 - y^2 = 1)$. This reflection is an isometry of the hyperbolic model on M^2 and, hence, this hyperbola is a geodesic of this model. Its projection from O to the disc $X^2 + Y^2 < 1$ is a chord, which is, thus, a geodesic of the Klein model.

Any plane section of the surface M^2, by a plane containing the origin, intersects M along a hyperbola that is a geodesic of the hyperbolic model.

Indeed, such a plane can be transformed to any other (and to the plane $x = 0$) by a linear transformation of the vector-space $\mathbb{R}^3$, which preserves the form f. It is an easy exercise in linear algebra to prove it, but we prefer to use a different trick.

Consider the symplectic plane $\mathbb{R}^2$ with symplectic structure $\omega = dp \wedge dq$ and Darboux linear coordinates p and q. The vector-space of the real quadratic forms on this plane is 3-dimensional. Writing the forms as

$$E\ p^2 + 2F\ pq + G\ q^2\ ,$$

we introduce the coordinates (E, F, G) in this 3-dimensional vector-space of quadratic forms on the symplectic plane.

The linear symplectomorphisms of the plane form the group $\mathrm{SL}(2, \mathbb{R})$, which is the group of determinant 1 matrices of linear transformations of the plane. The matrix of the quadratic form

$$\Phi = \begin{pmatrix} E & F \\ F & G \end{pmatrix}$$

is transformed by a linear change of variables $C \in \mathrm{SL}(2, \mathbb{R})$ to the matrix $C\Phi C^*$ (where C^* means the transpose matrix of C). The determinant of the form matrix is preserved:

$$\det(C\Phi C^*) = (\det C)(\det \Phi)(\det C^*) = \det \Phi\ ,$$

since $\det C = \det C^* = 1$.

Thus the group $\mathrm{SL}(2, \mathbb{R})$ acts on the vector-space $\mathbb{R}^3$ of quadratic forms on the symplectic plane, preserving the determinant $EG - F^2$ of the quadratic forms. This expression $EG - F^2$ can be reduced to the form $z^2 - x^2 - y^2$ by an easy linear change of variables. Thus we can interpret the action of the f-preserving linear transformations of the 3-space $\mathbb{R}^3$ as the natural action of the linear symplectomorphisms C on the space of the quadratic forms Φ defined on the symplectic plane.

From this point of view the solid cone $f > 0$ corresponds to the positive definite and to the negative definite forms Φ, while the complementary domain $f < 0$ represents the hyperbolic forms.

The surface M represents the determinant 1 forms. Thus the equivalence of the points of M with respect to the action of the linear transformations of the vector-space $\mathbb{R}^3$, preserving the form f, follows from the reducibility of any positive form of determinant 1 to the normal form $p^2 + q^2$ by a linear symplectomorphism.

Indeed, we first reduce it to the form $\alpha^2 p^2 + \frac{1}{\alpha^2} q^2$ by an orthogonal rotation, and then we use the symplectomorphism $P = \alpha p$, $Q = q/\alpha$.

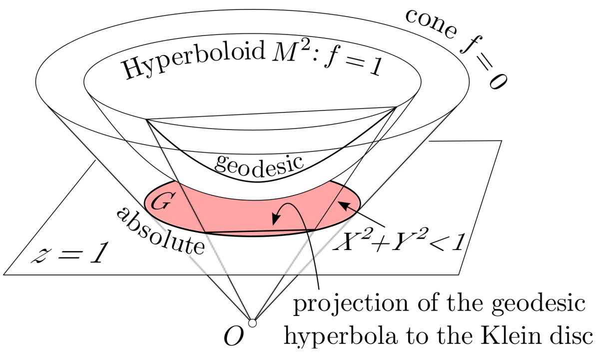
Thus, every point of M^2 may be reduced to any other (say, to the point on the z axis in the x, y, z coordinates). Then, by a rotation around this axis, we reduce any plane containing it, to any other such plane. This elementary theory of normal forms shows that all the sections of M^2 by the planes containing the origin, are reducible to the plane $x = 0$ by the f-preserving linear transformations of $\mathbb{R}^3$, and hence by the isometries of the surface M.

Therefore their projections to the disc $G : X^2 + Y^2 < 1$, which are just all its chords, are the geodesics of the Klein model of the Lobachevsky geometry.

There are no other geodesic lines, since at every point of the disc there exists a chord along every direction. Theorem 6 is proved. $\square$

The geometry of the chords evidently respects all the Euclidean geometry axioms: there exists one and only one chord connecting any two different points of the disc, and so on. The only exception is the Euclidean axiom on the parallel lines, which is evidently unsatisfied, while the Lobachevsky alternative axiom is evidently verified by the Klein model chords (Fig. 8.22), which do not intersect a given one.

The whole Lobachevsky plane geometry can be repeated in the Klein model, and it is rather convenient to represent the straight Lobachevsky lines by the straight Euclidean chords.



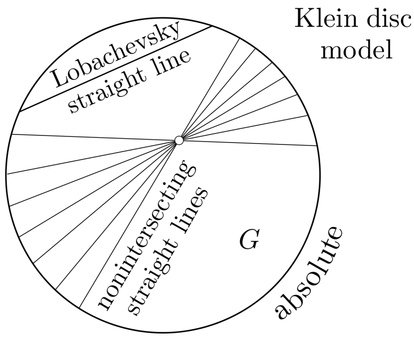


Figure 8.22: The projection of the hyperbolic model of the Lobachevsky plane to the Klein model disc (left) and its Lobachevsky axiom (right).

However, in opposite to the Poincaré model, the measures of the Lobachevsky angles differ from their representatives in the Klein model (say, orthogonality is not visible at the chord intersections), and circles, oricycles and equidistant curves are rather complicated objects in the Klein model.

The formula for the metric in the Klein model may be written explicitly in the coordinates X and Y, but it is rather long.

The Poincaré half-plane model and the Klein model of the Lobachevsky plane are isometric, and the isometric map may be presented geometrically. We leave this nice exercise to the readers. To solve it, it is useful to consider a model of the Lobachevsky plane that represents it as the interior of a circular disc, but whose geodesics are the circles orthogonal to the boundary of that disc. This model of the Lobachevsky plane is called the *Poincaré disc.* Notice that the image of a half-plane H under an inversion whose centre is not in H is the interior of a disc (to preserve the orientation one can make additionally a reflection preserving that disc).

There exists only one Lobachevsky plane up to isometries, but we shall not prove this theorem (whose proof is not too difficult), since it is purely negative, providing no new information on the real world around us.

8.8 The De Sitter world

The complement $\mathbb{R}\mathrm{P}^2 \setminus G$ to the disc G of the Klein model of the Lobachevsky plane is the Möbius band. The Klein model construction can be extended to this domain, but the resulting quadratic forms in the tangent spaces are not positive definite: Instead of a Riemannian metric, they define a *pseudo-Riemannian Lorenzian* (signature) "*metric*" on the Möbius band. This pseudo-Riemannian Lorenzian manifold is called *the relativistic De Sitter world.* We shall now describe it.

Instead of the locally convex two-sheeted hyperboloid M defined by the equation $f(x, y, z) = +1$ (Fig. 8.21, p. 299), we start from the one-sheeted hyperboloid N on which $f(x, y, z) = -1$.

The slightly perturbed version of this hyperboloid, $f = -1-\varepsilon$, intersects the affine tangent 2-planes of the smooth surface $N \subset \mathbb{R}^3$ along the hyperbolas (of "size" $\sqrt{\varepsilon}$). As we did for the Klein model, we divide by $\sqrt{\varepsilon}$ the tangent vectors to N (lying in the affine tangent plane) whose end-points belong to this hyperbola, and take the limits of the resulting tangent vectors for $\varepsilon \to 0$. In this way, we define in the tangent space a "size 1" hyperbola on which the quadratic form of signature $(+, -)$ is equal to 1.

Thus we define on the surface N a smooth field of Lorenzian hyperbolic quadratic forms on the tangent spaces, which is invariant under the action of the 3-dimensional Lie group formed by the linear transformations of the vector-space $\mathbb{R}^3$, that preserve f. It acts on the surface N as a group of diffeomorphisms.

Definition. The 2-dimensional *De Sitter relativistic world* is the 1-sheeted hyperboloid surface N, equipped with the pseudo-Riemannian (Lorenzian) metric that we have constructed above.

Remark. One might avoid our geometrical description, writing instead the coefficients of this quadratic form in some coordinates. It is in such way that the De Sitter world is usually introduced in the theoretical physics books.

Some simple geometric properties of this world (like the invariance under the action of Lie group $\mathrm{O}(2,1)$ of f-preserving linear transformations of the vector-space $\mathbb{R}^3$, or under the action of Lie group $\mathrm{SL}(2,\mathbb{R})$ of the linear symplectomorphisms of a symplectic plane) may be deduced by long calculations from the expressions of these components, provided that one will not make arithmetic mistakes.

We shall discuss the De Sitter world geometrically. This geometry is strangely missing in the physics textbooks.

Projecting the one-sheeted hyperboloid surface N to the real projective plane $\mathbb{R}\mathrm{P}^2$ (by the straight lines containing $O \in \mathbb{R}^3$), we cover twice the complementary Möbius band of the disc $G \subset \mathbb{R}\mathrm{P}^2$ of the Klein model. The two preimages of a point of the Möbius band are two opposite points of N, and the Möbius band is just the quotient space

$$N/(\mathbb{Z}_2 = \{1, -1\}) \ ,$$

which eliminates the doubling of the De Sitter world on the hyperboloid N.

We shall call this quotient space *the Möbius band model of the De Sitter 2-world.* On this model we get the projected Lorenzian pseudo-Riemannian metric (of signature $(+,-)$), and hence a field of crosses in the tangent spaces, formed by the zero lines of the quadratic forms (see Fig. 8.23). Their directions are called *asymptotic directions* in mathematics and *light directions* in physics (where the positivity and negativity directions of the quadratic form are called *time-like* and *space-like* directions).

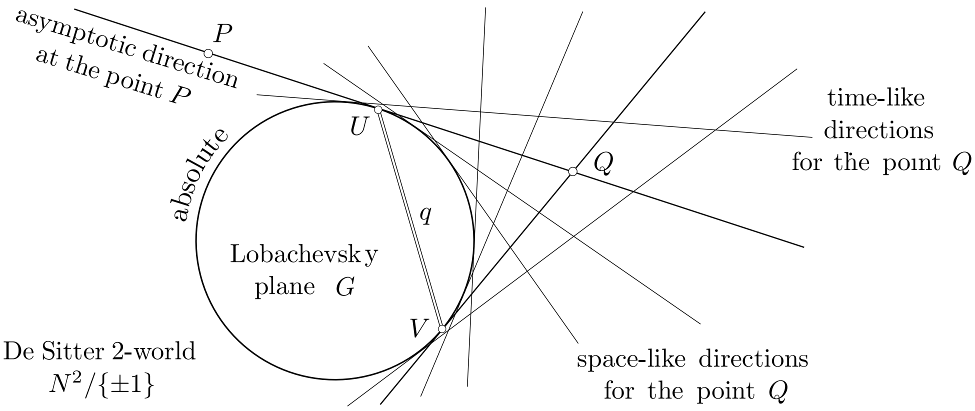


Figure 8.23: The Möbius band model of the De Sitter two-dimensional world.

The asymptotic directions on the surface of the hyperboloid N are, evidently, the directions of its straight lines, which are also the intersections of the hyperboloid N with its affine tangent planes in $\mathbb{R}^3$. Their projections to the Möbius band model of the De Sitter 2-world are the tangent lines to the absolute circle $X^2+Y^2=1$, separating the Klein model disc of Lobachevsky plane from the De Sitter exterior Möbius band (see Fig. 8.23).

In the interpretation in terms of quadratic forms (pp. 301-302), the Klein model disc represents the positive definite quadratic forms, its bounding absolute curve represents the degenerate quadratic forms, and the complementary Möbius band represents the hyperbolic forms (of signature $(+,-)$).

The 3-dimensional Lie group $\mathrm{SL}(2,\mathbb{R})$ of the symplectic transformations acts on the Möbius band model of the De Sitter relativistic 2-world by the diffeomorphisms preserving the Lorenzian structure. It can be also defined as the subgroup of those projective transformations of the plane $\mathbb{R}\mathrm{P}^2$ that preserve the Lobachevsky plane disc G and preserve, hence, the complementary Möbius band and the separating absolute circle.

The geodesic lines of the De Sitter world are represented in the Möbius band model by the straight lines of the real projective plane $\mathbb{R}\mathrm{P}^2$.

A point Q of the De Sitter world can be interpreted as a straight line q of the Lobachevsky plane: The two tangency points U and V of the asymptotic directions lines at a given point Q of the De Sitter world (Fig. 8.23) are connected inside the disc G of the Klein model of the Lobachevsky plane by a chord UV, which represents a straight line q in the Lobachevsky geometry.

Similarly, the points of the Klein model of the Lobachevsky plane can be interpreted as straight lines in the De Sitter 2-world Möbius band model. Namely, a Lobachevsky straight line that contains the given point represents a point of the De Sitter 2-world, and the set of all the Lobachevsky straight lines through the chosen point represents some curve of the De Sitter 2-world (which is a straight line of the De Sitter 2-world that has no intersection with the absolute curve).

In this way, the projective duality theory, defined by the absolute circular curve, provides a duality between the Lobachevsky plane (in the Klein model) and the De Sitter relativistic two-dimensional world (in the Möbius band model). This projective geometry has many applications, providing unexpected results both in the Lobachevsky geometry problems and in the relativistic geometry of the De Sitter 2-world*.

Example. The altitudes theorem for a Lobachevsky triangle is the geometric interpretation of the Jacobi identity in the Lie algebra of the 3-dimensional Lie group $\mathrm{SL}(2,\mathbb{R})$ of the linear symplectomorphisms of the symplectic plane $\mathbb{R}^2 = \{p, q\}$ (equipped with the symplectic form $\omega = dp \wedge dq$).

Namely, one interprets three elements u, v and w (of the space of quadratic forms on $\mathbb{R}^2$) as defining three sides of a triangle Δ in $\mathbb{R}\mathrm{P}^2$ (Fig. 8.24). The Poisson brackets $\{u, v\}$, $\{v, w\}$ and $\{w, u\}$ are then represented by the vertices of this triangle. We associate to a nonzero vector in the 3-space of the quadratic forms both a point W and a line w in $\mathbb{R}\mathrm{P}^2$, using the projective duality with respect to the "absolute circular curve" in $\mathbb{R}\mathrm{P}^2$, which represents the degenerate quadratic forms.

Some easy calculations show that, as a line, the product $\big\{\{u, v\}, w\big\}$ corresponds to the altitude of the triangle Δ, orthogonal to the w-side and containing the $\{u, v\}$ vertex. The Jacobi identity implies the intersection of the three altitude lines at a common point of $\mathbb{R}\mathrm{P}^2$, but this point may

*There are quite similar constructions for the n-dimensional De Sitter world and Lobachevsky space. One starts from the quadratic form $f = z^2 - x_1^2 - \cdots - x_n^2$ in $\mathbb{R}^{n+1}$.

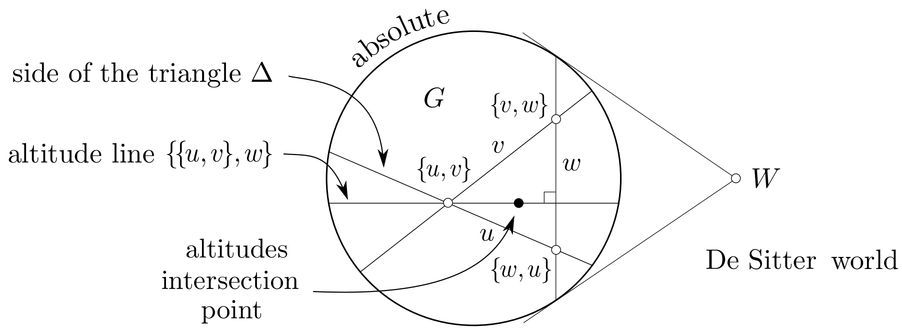


Figure 8.24: The altitudes theorem in Lobachevsky geometry as the Jacobi identity for the quadratic forms algebra.

belong either to the Klein model disc G of Lobachevsky plane, or to the complementary De Sitter 2-world Möbius band model. The last situation never happens, if all the Lobachevsky angles of the triangle Δ are smaller than $120°$.

In the example of Fig. 8.14, the altitudes of the triangle in the Poincaré model have no intersection point, even prolonging them to the lower half-plane. However, considering these lines in the Klein model, their intersection point lives in the relativistic De Sitter part of the projective plane $\mathbb{R}P^2$. This point is invisible in the Poincaré model representation of Fig. 8.14, since the De Sitter part of $\mathbb{R}P^2$ becomes complex in the Poincaré model.

The details on the theorems on the altitudes in Lobachevsky and De Sitter geometries are given in [35].

8.9 Modular coverings of Lobachevsky plane

Many geometrical objects look very differently in the Poincaré model representation and in the Klein model representation of the Lobachevsky geometry.

The reflection group of the infinite Lobachevsky triangle, with three vertices on the absolute curve, is represented in Fig. 8.25. The reflections in the sides of such a triangle generate an (infinite) Lobachevsky reflections group. The images of the initial (infinite) equilateral triangle under these reflections are shown in Fig. 8.25.

In the Klein model representation, the images are still Euclidean triangles that approach the absolute curve closer and closer, when the number of reflections increases.

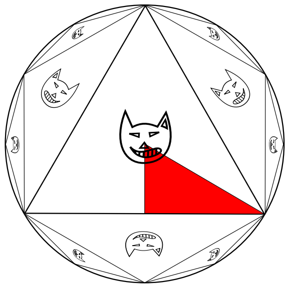


Figure 8.25: The "modular covering" of the Lobachevsky plane by the reflected triangles in the Klein model representations.

Definition. A *fundamental domain* of the action of a symmetry group on a given space E (for instance, on the upper half-plane), is a connected region of E that contains exactly one representative of each orbit.

The symmetries of the covering of the Lobachevsky plane by this infinite set of triangles, include also the reflections in the medians of these triangles. A fundamental domain (covering 1/6 of the initial triangle) is shown in Fig. 8.25: It is a Lobachevsky triangle with angles $\pi/2$, $\pi/3$ and 0. This symmetry group is a discrete subgroup of the Lie group of the isometries of the Lobachevsky plane: The points of its orbits are isolated and have no other points of the same orbit in some neighbourhood.

In the Poincaré half-plane model (Fig. 8.26), the triangles are bounded by the circles orthogonal to the absolute curve. To show the similarity to the Klein model features on Fig. 8.25, we have sent the Poincaré model half-plane to the interior of a Poincaré disc, by the usual complex projective transformation.

The modular group $\mathrm{SL}(2,\mathbb{Z})$ is the subgroup (of motions, without reflections) of index 2 of the symmetry group described above. This becomes clear if we write down the action in the Poincaré model. For example, the symmetry shift $z \mapsto z+1$, observed in Fig. 8.26, is realised by the linear-fractional transformation whose matrix is $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. The matrix $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ is also realised and these two transformations generate the whole group $\mathrm{SL}(2,\mathbb{Z})$.

It would be interesting to understand whether the action of the modular group $\mathrm{SL}(2,\mathbb{Z})$ on the De Sitter 2-dimensional world (of the Möbius band model) is discrete: It consists of the same projective transformations that act discretely on the points of the Lobachevsky plane (as shown in Fig. 8.25 and in Fig. 8.26). But it is not clear whether this action is still discrete on

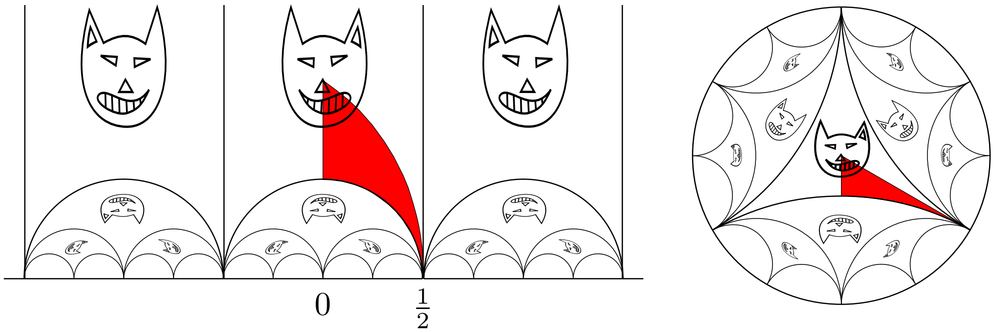


Figure 8.26: The "modular covering" of the Lobachevsky plane by the reflected triangles in the Poincaré model representations.

the De Sitter world of the Lobachevsky straight lines (these might even be a Lobachevsky straight line whose images under the modular group action approximate every Lobachevsky line in the De Sitter world topology?). The union of the reflection mirrors in Fig. 8.25 is everywhere dense in the De Sitter part of the projective plane $\mathbb{RP}^2$, and these mirrors belong to the orbit of one of the sides of the initial triangle.

The fundamental domain of the modular group can be obtained, for example, as the union of the triangle (shaded on Fig. 8.26) that forms the fundamental domain of the whole symmetry group with the neighbouring symmetric triangle, so that the union is a triangle with angles $(0, 0, 2\pi/3)$ and whose vertices having angle 0 are lying on the absolute.

This follows from the fact that the symmetry $z \mapsto -\overline{z}$ is an isometry (a reflection) of the Poincaré model of the Lobachevsky plane to itself that reverse the orientation.

It is possible also to take other neighbouring triangles. For example, it is possible to obtain a fundamental triangle of the modular group with angles $(0, \pi/3, \pi/3)$.

To study the orbits of the modular group $\mathrm{SL}(2, \mathbb{Z})$ on the De Sitter 2-world, we shall use the following property of the geodesic lines in the quotient space of the Poincaré model of the Lobachevsky plane by the natural action of the modular group.

We can represent this quotient space by the fundamental domain, which is an infinite triangle of finite area. In Fig. 8.27, we show the fundamental triangle ABC of the corresponding reflection group.

A geodesic line, represented in the Poincaré model as a circle orthogonal

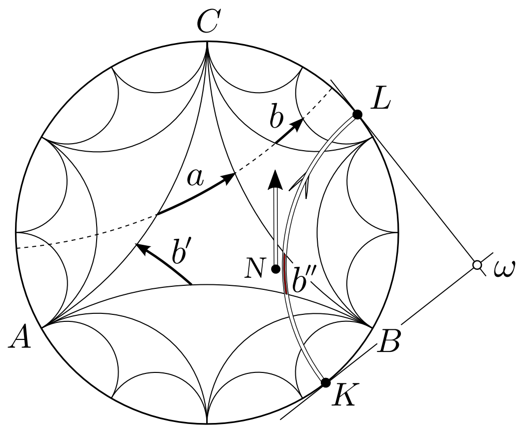


Figure 8.27: The images b' and b'' of the segments of the continuation of the geodesic segment a in the modular triangle ABC.

to the absolute, is (generically) leaving a fundamental domain, to come to a neighbouring one, and so on – Fig. 8.27.

Showing these next parts of the geodesic line in the initial fundamental domain, we obtain there a sequence of geodesic segments $a, b', \ldots$: Each of these segments b' is the image of the corresponding segment b of the initial geodesic line, transported to the initial fundamental domain by an element of the group $\mathrm{SL}(2, \mathbb{Z})$.

In ergodic theory of dynamical systems it is proved that this geodesic flow is ergodic* (in the space of the unit tangent vectors of the quotient space).

It follows that a typical geodesic line is dense in this space. In other words, for any point N of the fundamental domain and for any direction at this point there exist a transported segment b'' that crosses the fundamental domain approximating there the given point N and the given direction, arbitrarily close (Fig.8.27).

---

*Roughly speaking, *ergodic* means that it has good mixing properties. There are at least two possible ways to formulate this. Let $\{g_t : X \to X\}$ be a family of diffeomorphisms preserving the volumes.
Definition 1: The flow g_t is *ergodic* if for any (measurable) subset A in X its image $g_t(A)$ is 'spread homogeneously' along X, for big t. More explicitly, the volume of $g_t(A) \cap B$ is proportional to the volumes of A and B and is independent of t when t tends to infinity.
Definition 2: The flow g_t is *ergodic* if its trajectories $g_t(x), t \in \mathbb{R}$, are 'spread homogeneously' along X, at least for generic x. More explicitly, the average time t for which $g_t(x)$ belongs to B is proportional to the volume of B. Under certain conditions these two definitions are equivalent

Now, we interpret each geodesic line KL of the Lobachevsky plane as the intersection point ω of the two tangent lines to the absolute circle at the end-points K, L of that geodesic. Of course, this intersection point ω belongs to the De Sitter world. We have proved the

Theorem 7. *Most orbits of the natural action of the modular group on the De Sitter world (Möbius band) are everywhere dense subsets of this De Sitter manifold.*

There exist also exceptional non dense orbits, forming a set of measure zero (say, those corresponding to the closed geodesic lines are finite).

Chapter 9

Riemannian geometry

A *Riemannian metric* on a smooth manifold M is a family of positively definite quadratic forms in the tangent spaces depending smoothly on the points of the manifold. In Ch. 8, we have considered the theory of 2-dimensional Riemannian manifolds (parallel transport, geodesic curvature, Gauss curvature, etc.), focusing on Lobachevsky plane geometry. In this chapter, we shall do it for higher dimensional Riemannian manifolds.

The "inner" (or "interior") properties of Riemannian manifolds are those determined by the Riemannian metric (persisting under any isometry). For example, when bending (i.e., mapping isometrically) a sheet of paper into a cone or a cylinder, the lengths, angles and areas inside it are not changed, being inner features of Riemannian surfaces.

The invariant that distinguishes Riemannian metrics (generalising Gaussian curvature) is called *Riemannian curvature*. It encodes the local behaviour of geodesics on the manifold (their local rapprochement or repulsion). For example, if the curvature is negative the neighbouring geodesics rapidly diverge from one another, as on Lobachevsky plane.

9.1 Geometry of smooth hypersurfaces

A hypersurface is a submanifold of codimension 1, that is, of dimension $n-1$ in a manifold of dimension n.

Example. A point is a hypersurface of a curve. The hypersurfaces of a 2-dimensional manifold are the curves. A plane in $\mathbb{R}^3$ is a hypersurface.

Submanifolds of Riemannian manifolds inherit the Riemannian metric. The induced metric is called *first fundamental quadratic form*. In this section, we shall consider hypersurfaces in Euclidean space $\mathbb{R}^{n=m+1}$.

Example. The simplest example is provided by a surface in Euclidean space $\mathbb{R}^3$. If the surface is given parametrically as $r(u_1, u_2) \in \mathbb{R}^3$, then the coefficients of the metric $g = Adu_1^2 + 2Bdu_1du_2 + Cdu_2^2$ ($g_{11} = A$, $g_{12} = g_{21} = B$, $g_{22} = C$) are equal to the scalar products of the basic tangent vectors

$$g_{ij} = \langle\, r_i, r_j \,\rangle\,, \quad \text{with} \;\; r_1 = \frac{\partial r}{\partial u_1}\,,\;\; r_2 = \frac{\partial r}{\partial u_2}\,.$$

Exercise. Verify that the above coefficients g_{ij} of the metric g are obtained from the ambient Euclidean metric $dx^2 + dy^2 + dz^2$ by substituting the differentials of the component functions of the parametrised surface $r(u_1, u_2) = (x(u_1, u_2), y(u_1, u_2), z(u_1, u_2))$, written in terms of the parameters u_1, u_2: $dx = \frac{\partial x}{\partial u_1}du_1 + \frac{\partial x}{\partial u_2}du_2$, and so on.

Pseudo-example. The form $dx^2 + dy^2 - dz^2$ in $\mathbb{R}^3$ is not a metric because it is not positively definite. However, as we have seen in Ch. 8 (p. 299), its restriction to the two-sheets hyperboloid $z^2 - x^2 - y^2 = 1$ provides the metric of the hyperbolic model of Lobachevsky plane (also called *pseudosphere*).

The inner geometry of a hypersurface of $\mathbb{R}^n$ is determined by its first quadratic form (its induced metric), while its outer (or "exterior") geometry is determined by the way in which it is embedded in $\mathbb{R}^n$. We shall now consider two important outer invariants that characterise how the hypersurface is placed in ambient space:

Gauss map and second fundamental form. The *Gauss map* $\Gamma : M \to \mathbb{S}^m$ of a co-oriented hypersurface M in Euclidean space $\mathbb{R}^{n=m+1}$ is defined as follows. For each point $x \in M$ one translates parallelly the co-orienting normal unit vector $\omega(x)$ to the origin to get a vector $\Gamma(x) \in \mathbb{S}^m \subset \mathbb{R}^{m+1}$. Since both tangent hyperplanes T_xM and $T_{\Gamma(x)}\mathbb{S}^m$ are orthogonal to $\omega(x)$, their natural identification allows to consider the Gauss map derivative Γ_* as a linear transformation of $T_xM \approx \mathbb{R}^m$. We shall now compute the matrix associated to this linear map Γ_* (the Jacobian matrix of Γ) and its determinant.

Suppose we study the hypersurface provided by the graph a smooth function of m variables (we can always suppose it locally), and that the independent coordinates $(x_1, \dots, x_m)$ on the horizontal m-dimensional plane vanish at the point we are studying, as well as the vertical coordinate y:

$$M = \{(x, y) : y = f(x)\}, \quad f(0) = 0, \quad df(0) = 0.$$

We choose the co-orienting normal vector $\nu(x)$ given by the coordinates

$$\dot{x}_1 = \partial f/\partial x_1, \dots,\; \dot{x}_m = \partial f/\partial x_m,\; \dot{y} = -1.$$

(we can take the gradient of the function of $m+1$ variables $f(x)-y$). To write the coordinates of the unit normal vector ω (providing locally the Gauss map of M) we normalise the vector ν to obtain $\Gamma(x)=\omega=\nu/|\nu|\in\mathbb{S}^m$.

Now, to find the Jacobi matrix of Γ at $x=0$, $A=(\partial\omega_i/\partial x_j)$, we use the evident relation $|\nu(x)|=1+O(|x|^2)$ which implies that the division by $|\nu(x)|$ does not change the derivatives at $x=0$:

$$\frac{\partial\omega_i}{\partial x_j}|_0=\frac{\partial\nu_i}{\partial x_j}|_0=\frac{\partial^2 f}{\partial x_i\partial x_j}|_0.$$

Therefore the Jacobi matrix of the Gauss map Γ of M at $x=0$, which represents its derivative Γ_*, is equal to the symmetric matrix of the second partial derivatives of f, $||f_{x_ix_j}||$. It is called the *Hessian matrix of f* (we shall denote it by $||h_{ij}||$, with $h_{ij}:=f_{x_ix_j}$) and its determinant is called *the Hessian of f*. To compute it, it suffices to take the Taylor series of f at the point $x=0$, which starts with the quadratic term,

$$f(x)=\frac{1}{2}\sum f_{x_ix_j|_0}x_ix_j+o(|x|^2).$$

This quadratic part, $h(x):=\frac{1}{2}\sum f_{x_ix_j}\,x_ix_j$, is called *the second fundamental form* of the co-oriented hypersurface M at the considered point ($x=0$ in our coordinate system); its associated linear symmetric operator $Q:\mathbb{R}^m\to\mathbb{R}^m$, $h(x)=\frac{1}{2}\langle Qx,x\rangle$, coincides with Γ_*. The equality $f(x)=\frac{1}{2}\langle\, Qx\,,\,x\,\rangle+o(|x|^2)$ expresses that the second fundamental form measures the quadratic deviation of the surface from its affine tangent plane at the considered point. The co-orientation is defined by the choice of the normal vector ν.

Remark. Since h is applied to vectors of the horizontal hyperplane tangent to M at $x=0$, it is often convenient to consider it as a form defined on the "abstract" tangent hyperplane T_0M and to write it using differentials $h=\frac{1}{2}f_{x_ix_j}\,dx_idx_j$ (for $m=2$ we have $h=f_{xx}dx^2+2f_{xy}dxdy+f_{yy}dy^2$).

Remark. We can also see h as a symmetric bilinear form. It is just the second differential $d^2f_{x=0}$, which is well defined and does not depend on the coordinate system used to write the Taylor series because our point $x=0$ is critical, $df_{x=0}=0$; its value on a pair of tangent vectors ξ, η is provided by the mixed directional derivatives $h(\xi,\eta)=\partial_\xi\partial_\eta f(0)$ (verify it!).

Principal curvatures, principal directions and Gauss curvature. A basic result of linear algebra states that all eigenvalues of a quadratic form are real and the eigenvectors of distinct eigenvalues are orthogonal *. Hence choosing m eigendirections as coordinate axes in the m-dimensional Euclidean tangent space to M at our point, we reduce the second quadratic form to the simple diagonal expression,

$$\frac{1}{2}\langle Qx\,,\,x\rangle = \sum_{j=1}^{m}\frac{1}{2}k_j x_j^2.$$

Definition. The *principal curvatures* of the hypersurface at the point $x = 0$ are the eigenvalues $k_i \in \mathbb{R}$ of the linear symmetric operator $Q : \mathbb{R}^m \to \mathbb{R}^m$ (the signs of the principal curvatures change if the opposite co-orientation is used); the *principal directions* are the corresponding eigendirections (they are defined unambiguously if all the m principal curvatures are different) and the *Gaussian curvature* is the product $K = k_1 k_2 \cdots k_m = \det\left(f_{x_i x_j}\,|_0\right)$ of the principal curvatures (the determinant of the Hessian matrix of f).

The factor $1/2$ in the diagonal expression of the second quadratic form was put there just to have these simple expressions for the curvatures.

Problem. Find the principal curvatures and the Gaussian curvature of the m-dimensional sphere of radius R in Euclidean n-space ($n = m + 1$).

Answer. The principal curvatures are $k_j = 1/R$, and hence the Gaussian curvature is $K = 1/R^m$ (so for the circle in the plane it is $1/R$ and for the sphere in $\mathbb{R}^3$ it is $1/R^2$).

Case of curves. Consider a plane curve represented in the neighbourhood of a point as the graph of a smooth function $f(x) = \frac{1}{2}kx^2 + \ldots$. In this case, $m = 1$, the single curvature k is the inverse value of the radius of the *osculating circle* of the curve at that point (it is the circle that better approximates the curve, than any other circle, at the considered point).

Parametrising a curve γ by its length s, starting from some of its points, and denoting by φ the angle of the unit tangent vector, $d\gamma/ds$, with a fixed direction, the curvature of γ at $s = s_0$ is equal to $k = \frac{d\varphi}{ds}(s_0)$.

Problem. Draw the osculating circle for a given generic point of a given generic curve.

Hint. Leibniz gave the wrong answer to this problem, shown in Fig. 9.1, in his first course of calculus. Newton would never committed this mistake, knowing that at the point under

---

*Indeed, if $Q : \mathbb{R}^m \to \mathbb{R}^m$ is a symmetric operator (i.e., $\langle Q\xi, \eta\rangle - \langle \xi, Q\eta\rangle = 0$ for any $\xi, \eta \in \mathbb{R}^m$) and there exist $\xi, \eta \in \mathbb{R}^m$, $\lambda, \mu \in \mathbb{R}$ such that $Q\xi = \lambda\xi$ and $Q\eta = \mu\eta$ with $\lambda \neq \mu$, then $0 = \langle Q\xi, \eta\rangle - \langle \xi, Q\eta\rangle = \langle \lambda\xi, \eta\rangle - \langle \xi, \mu\eta\rangle = (\lambda - \mu)\langle \xi, \eta\rangle$, that is, $\langle \xi, \eta\rangle = 0$.

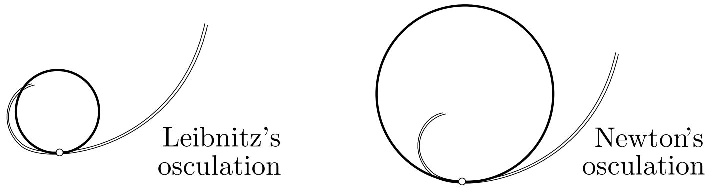


Figure 9.1: Leibniz and Newton osculating circles of a curve in Euclidean plane.

study the osculating circle has three (infinitely close) intersection points with the curve (similarly to the two infinitely close intersection points with the tangent straight line). Painters like Leonardo da Vinci would never make Leibniz mistake.

Since the integer 3 is odd, the osculating circle cross the curve tangentially, going from one side of the curve to the other (at a generic point of a generic curve).

Returning to the Gaussian curvature, we get the

Corollary. *The Gaussian curvature of a smooth surface in Euclidean* 3*-space is positive if the second quadratic form is positive definite or negative definite* (at a minimum point or at a maximum point of the function f)*, being negative at the points where the second quadratic form is hyperbolic* (corresponding to the saddle points of f, see Fig. 9.2).

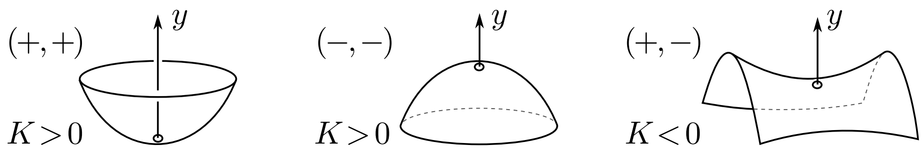


Figure 9.2: Signs of the Gaussian curvature of the surfaces $y = f(x_1, x_2)$.

Remark. To define the Gauss curvature and its sign we don't need to orient the (hyper)surface: It suffices to orient the parallel tangent hyperplanes T_xM and $T_\omega\mathbb{S}^m$ (where $\omega = \Gamma(x)$) both in the same way. If $m = 2$ the sign is explicitly described by the above corollary. For arbitrary m the sign is "+" if the number of negative principal curvatures is even, being "−" if it is odd.

The Gaussian curvature K measures the ratio of the oriented m-volumes: $d\omega = Kds$ ($d\omega$ = volume on $\mathbb{S}^m$, ds = volume on M).

Asymptotic curves and godrons of a surface

On a generic surface in Euclidean 3-space (Fig. 9.3) the points with negative Gaussian curvature, called *hyperbolic points* (where the tangent plane intersects the surface along two smooth intersecting curves), are separated from the points with positive Gaussian curvature, called *elliptic points* (where the surface is locally convex, having locally no common points with its tangent plane, except the tangency point), by a smooth *parabolic curve* (where the second quadratic form $h = f_{xx}dx^2 + 2f_{xy}dxdy + f_{yy}dy^2$ is degenerated).

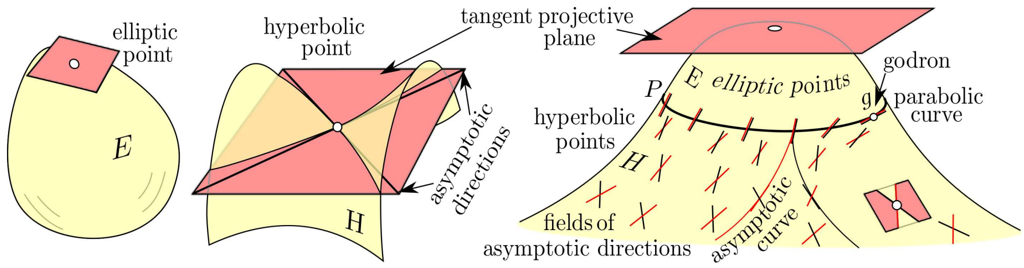


Figure 9.3: Elliptic, parabolic and hyperbolic points of a surface in the projective 3-space. The hyperbolic domain has two fields of asymptotic straight lines.

On the hyperbolic domain (H) we get two fields of *asymptotic straight lines*: The tangent lines to the curves of intersection of the surface with its tangent plane at the considered point (at which they have at least 3-point contact with the surface). If the surface is locally the graph of a smooth function $z = f(x, y)$ the asymptotic directions are given by the equation:

$$f_{xx}dx^2 + 2f_{xy}dxdy + f_{yy}dy^2 = 0 \qquad (1)$$

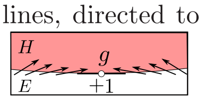
The integral curves of these fields of lines are called *asymptotic curves* (Fig. 9.3).

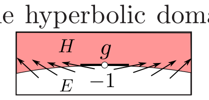
Exercise. Find the asymptotic curves on the hyperboloid surface of one sheet.

Answer. They are just the generating straight lines of the hyperboloid (Fig. 3.12, p. 69).

On the parabolic curve (P) there is a unique, but double, asymptotic straight line. On the elliptic domain (E) there is no real asymptotic line (being complex conjugate). So the parabolic curve is provided by the discriminant curve of (1): $f_{xy}^2 - f_{xx}f_{yy} = 0$.

A *godron* (or *cusp of Gauss*) is a parabolic point at which the single asymptotic line is tangent to the parabolic curve (Fig. 9.3). Godrons have many interesting properties; for example, *any smooth curve of a surface of* $\mathbb{R}^3$ *tangent to the parabolic curve at a godron* g *has at least* 4*-point contact with the tangent plane of the surface at* g.

The local behaviour of the asymptotic lines of generic surfaces separates godrons into *positives* and *negatives*: A godron has *index* $+1$ (resp. *index* -1) if at its neighbouring parabolic points the half-asymptotic lines, directed to the hyperbolic domain, point towards (resp. away from) the godron: 



.

To each godron g of a generic surface S is associated a number $\rho(g) \neq 1$ which is a projective invariant (defined in [126]). Near g, one can always put the 4-jet of S (after projective transformations) into the so-called "Platonova's normal form":

$$z = \frac{y^2}{2} - x^2 y + \frac{\rho}{2} x^4. \tag{2}$$

Exercise. Let S be the surface $z = z(x, y)$ given by equation (2).
1. In the xy-plane, draw the "Hessian curve", that is, draw the curve corresponding to the parabolic curve of S. Then indicate the elliptic and hyperbolic domains.
2. Use (2) to prove that a godron is positive (negative) if and only if $\rho > 1$ (resp. $\rho < 1$).
3. For which values of ρ the local intersection of S with its tangent plane at g consist of g ?

Remark. Unlike the curvatures and the principal directions, the asymptotic curves and the godrons are independent of any Euclidean or Riemannian metric (needed to define the curvatures): They are well defined in the projective case. Indeed, in the projective space there is a 2-dimensional projective subspace tangent to the surface at the given point and intersecting the surface along two smooth curves whose tangent directions define the asymptotic directions of the surface at the tangency point.

In thermodynamics, the godrons of the graph of the "internal energy function" of a fluid are crucial to describe its phase changes (see p. 613). This led Korteweg to study the properties of smooth surfaces based on their tangency with straight lines. For example, he studied the bifurcations of the asymptotic, parabolic and "conodal" curves, but his works were unknown or forgotten (see [96]). One hundred years later, different research teams in Singularity Theory have carried out again the study of "tangential" singularities and of bifurcations of the parabolic curve (always involving godrons). The book [44] is devoted to godrons; more recent results are given in [126].

Problem. Find the global behaviour of the asymptotic curves of a generic cubic surface in $\mathbb{RP}^3$ defined by a generic homogeneous polynomial equation of degree 3: $f(x, y, z, t) = 0$, say, $x^3 + y^3 + z^3 + t^3 = axyz$ for a generic constant a.

The question is to understand the global topological properties of the asymptotic curves. Are they covering the hyperbolic domain densely? The conjecture is that they behave chaotically and that the corresponding differential equation has no analytic (or even smooth) first integrals, like in the three-body problem of celestial mechanics. But it is unproved, and no one has practically made the picture of the asymptotic curves of a generic cubic surface neither in affine nor in projective 3-space (neither real nor complex).

More properties of asymptotic curves are considered in Ch. 19, "Fair Problems".

9.2 Theorema Egregium

The German mathematician Carl Friedrich Gauß [Gauss (1777-1855)] discovered the existence of an "inner" geometry of manifolds. In his "Theorema

Egregium" (1827), on 2-dimensional manifolds (i.e., surfaces), he arrived to the concept of "curvature" at a point, $1/R_1R_2$ (where $R_1 = 1/k_1$, $R_2 = 1/k_2$ are the principal radii of curvature of the surface at the point), stating that it is an inner invariant. Later, his student B. Riemann extended the theory to higher dimensional manifolds, as we shall see below.

Gauss' personal work on practical problems was a very important motivation for his investigations in "pure" mathematics. Indeed, Gauss worked in many applied areas from which we mention some of them : Astronomy (he developed tools for the study of perturbations, he found criteria of convergence and from 1807 till his death Gauss was director of the Goettingen observatory), Physics (he made contributions in electricity, telegraphy, mechanics, potential theory, theory of capillary action, terrestrial magnetism, etc.), Geodesy (his first problem was to survey the Kingdom of Hanover and, doing it, his practical land surveys lead him to invent, in 1821, the "heliotrope" [an instrument that uses a mirror to reflect sunlight over great distances to mark the positions of participants in a land survey]), etc. Concerning geodesy, Gauss knew that a flattened ellipsoid of rotation is an approximate representation of the figure of earth and, looking for a theory to elaborate precise geographical charts, he studied the deviation of the true earth shape from that ellipsoid. During those years of practical activity Gauss published his "Method of least squares" and his "Disquisitiones generales circa superficies curvas", including his Theorema Egregium.

We present Theorema Egregium based on the concept of "connection 1-form", as was suggested to us by M. Kazarian. It provides simple formulas to compute the Gaussian curvature in terms of the Riemannian metric.

Connection 1-form. Given a two-dimensional Riemannian manifold, we start "orthogonalising" the metric by expressing it with an appropriate orthonormal basis of 1-forms $u_1, u_2 \in \Omega^1$, that is, a base dual to an orthonormal basis e_1, e_2 of tangent vectors :

$$g = g_{11}dx^2 + 2g_{12}dxdy + g_{22}dy^2 = u_1^2 + u_2^2.$$

With these 1-forms u_1, u_2 the area 2-form is expressed by $\sigma = u_1 \wedge u_2$.

Moreover, since any 2-form is a multiple of σ, the exterior derivatives $du_1 = \alpha_1\sigma$, $du_2 = \alpha_2\sigma$ determine two functions α_1, α_2 that we use in the

Definition. The *connection* 1-*form* (determined by u_1, u_2) is defined as

$$\alpha = \alpha_1 u_1 + \alpha_2 u_2 \in \Omega^1 .$$

Exercise. Prove that the area form is independent of the chosen basis u_1, u_2.

Hint. Any other orthonormal basis $\widetilde{u}_1, \widetilde{u}_2$ is obtained by a rotation of u_1, u_2,

$$\widetilde{u}_1 = \cos\psi u_1 + \sin\psi u_2, \quad \widetilde{u}_2 = -\sin\psi u_1 + \cos\psi u_2 .$$

Thus the area form is invariant: $\widetilde{u}_1 \wedge \widetilde{u}_2 = u_1 \wedge u_2 = \sigma$.

Lemma (proved below). *Consider another orthonormal basis $\widetilde{u}_1, \widetilde{u}_2$. The associated connection form $\widetilde{\alpha}$ differs from α by the differential of a function: $\widetilde{\alpha} = \alpha + d\psi$.*

Since the derivative of α is a 2-form, it is a multiple of σ: $d\alpha = -K\sigma$.

Curvature form. The function K is called the *curvature* of the Riemannian surface (M, g). The 2-form $-d\alpha = K\sigma$ is called the *curvature form*.

Theorem (Gauss' Theorem). *The following holds:*
(i) *The curvature K is uniquely determined by the Riemannian metric and is independent of the choice of u_1 and u_2;*
(ii) *If the Riemannian metric is the first quadratic form of a surface $M \subset \mathbb{R}^3$, then K coincides with the Gaussian curvature of M.*

Proof of part (i). From the lemma we get $d\widetilde{\alpha} = d\alpha = -K\sigma$. □

Before to prove part (ii) and the lemma, we shall use Gauss' theorem:

Problem. Calculate the curvature of the following two metrics:
(a) $g = dx^2 + x^2dy^2$; (b) $g = dx^2 + 2\cos\omega(x,y)dxdy + dy^2$.

Answer. (a) $K = 0$; (b) $K = -\omega_{xy}/\sin\omega$.

Solution. (a) The equality $dx^2 + x^2dy^2 = dx^2 + (xdy)^2$ implies that

$$u_1 = dx, \quad u_2 = xdy \quad \text{and} \quad \sigma = xdx \wedge dy.$$

Since, $du_1 = 0$ implies $\alpha_1 = 0$ and $du_2 = dx \wedge dy$ implies $\alpha_2 = 1/x$, we get $\alpha = 0 \cdot dx + (1/x)xdy = dy$. That is, $d\alpha = 0$ and hence $K = 0$.

(b) From $dx^2 + 2\cos\omega\, dxdy + dy^2 = (dx + \cos\omega\, dy)^2 + (\sin\omega\, dy)^2$ we get

$$u_1 = dx + \cos\omega\, dy, \quad u_2 = \sin\omega\, dy \quad \text{and} \quad \sigma = \sin\omega\, dx \wedge dy.$$

In this case, the equality $du_1 = -\omega_x \sin\omega\, dx \wedge dy$ implies $\alpha_1 = -\omega_x$ and the equality $du_2 = \omega_x \cos\omega\, dx \wedge dy$ implies $\alpha_2 = \omega_x \cot\omega$. Consequently

$$\alpha = -\omega_x(dx + \cos\omega\, dy) + \omega_x \cot\omega\, \sin\omega\, dy = -\omega_x dx,$$

that is, $d\alpha = \omega_{xy}dx \wedge dy$ and hence $K = -\omega_{xy}/\sin\omega$.

Corollary (of the Lemma)**.** *A Riemannian metric g is Euclidean* (reduces to $dx^2 + dy^2$ in some local coordinates) *if and only if* $K \equiv 0$.

Proof. From $K \equiv 0$ we get $d\alpha = 0$ which implies the existence of a function ψ such that $\alpha + d\psi = 0$. The orthonormal basis $\widetilde{u}_1, \widetilde{u}_2$ associated to ψ determines a connection 1-form $\widetilde{\alpha}$ that satisfies $\widetilde{\alpha} = \alpha + d\psi = 0$ implying that $d\widetilde{u}_1 = d\widetilde{u}_2 = 0$. Hence $\widetilde{u}_1 = dX, \widetilde{u}_2 = dY$, that is, $g = dX^2 + dY^2$. $\square$

Problem (continuation of the previous problem).
(a) Since $K = 0$, find the functions X, Y for which $g = dX^2 + dY^2$.
(b) Find the conditions on the function ω in order to have a Euclidean metric and, then, find the functions X, Y for which $g = dX^2 + dY^2$.

Solution. (a) Since $\alpha = dy$, we have $\psi = -y$. Hence

$$\begin{aligned}
\widetilde{u}_1 &= \cos(-y)dx + \sin(-y)xdy = d(x\cos y),\\
\widetilde{u}_2 &= -\sin(-y)dx + \cos(-y)xdy = d(x\sin y).
\end{aligned}$$

Thus $X = x\cos y,\ Y = x\sin y$, and hence $dX^2 + dY^2 = dx^2 + x^2dy^2$. With the "change of variables" $r = x$, $\vartheta = y$, one sees that $g = dx^2 + x^2dy^2$ is the expression of the Euclidean metric in polar coordinates.

(b) The Euclidean metric condition $K = 0$ implies $\omega_{xy} = 0$, which means that $\omega(x, y) = a(x) + b(y)$. Thus $\alpha = -\omega_x dx = -a'(x)dx = -da(x)$, that is, $\psi = a(x)$. From this we get the equations

$$\begin{aligned}
\widetilde{u}_1 &= \cos a(dx + \cos(a+b)dy) + \sin a \sin(a+b)dy\\
&= \cos a(x)dx + \cos b(y)dy = d\left(\int \cos a(x)dx + \int \cos b(y)dy\right),\\
\widetilde{u}_2 &= -\sin a(dx + \cos(a+b)dy) + \cos a \sin(a+b)dy\\
&= -\sin a(x)dx + \sin b(y)dy = d\left(\int \sin a(x)dx + \int \sin b(y)dy\right).
\end{aligned}$$

Consequently $dX = \cos a(x)\,dx + \cos b(y)\,dy$, $dY = \sin a(x)\,dx + \sin b(y)\,dy$. The reader can verify that $dX^2 + dY^2 = dx^2 + 2\cos(a+b)dxdy + dy^2$.

2nd (equivalent) definition. The *connection* 1-*form* α, determined by an orthonormal basis of 1-forms u_1, u_2, is defined by the following equalities

$$du_1 = \alpha \wedge u_2\,, \quad du_2 = -\alpha \wedge u_1\,.$$

Proof of the equivalence. From the first definition of the connection 1-form we get the following equalities:

$$\begin{aligned}\alpha \wedge u_2 &= (\alpha_1 u_1 + \alpha_2 u_2) \wedge u_2 = \ \ \alpha_1 \sigma \ = \ du_1, \\ \alpha \wedge u_1 &= (\alpha_1 u_1 + \alpha_2 u_2) \wedge u_1 = -\alpha_2 \sigma = -du_2,\end{aligned}$$

That is, $\alpha \wedge u_2 = du_1$, $\alpha \wedge u_1 = -du_2$. The converse is also direct. □

Proof of the Lemma. Deriving the equation $\widetilde{u}_1 = \cos\psi \cdot u_1 + \sin\psi \cdot u_2$ and using the second definition of the connection 1-form we get the equalities

$$\begin{aligned} d\widetilde{u}_1 &= d\psi \wedge (-\sin\psi \cdot u_1 + \cos\psi \cdot u_2) + \cos\psi \cdot \alpha \wedge u_2 - \sin\psi \cdot \alpha \wedge u_1 \\ &= d\psi \wedge (-\sin\psi \cdot u_1 + \cos\psi \cdot u_2) + \alpha \wedge (-\sin\psi \cdot u_1 + \cos\psi \cdot u_2) \\ &= (d\psi + \alpha) \wedge \widetilde{u}_2 \end{aligned}$$

In the same way, we obtain the equality $d\widetilde{u}_2 = -(d\psi + \alpha) \wedge \widetilde{u}_1$. Using the second definition, these equalities imply that $\widetilde{\alpha} = d\psi + \alpha$. □

Proof of Gauss' Theorem *(ii)*.

Write e_1, e_2 for the tangent vectors dual to the basis u_1, u_2 and construct the orthonormal basis e_1, e_2, $n = e_3 = e_1 \times e_2$ at each point of the surface. For a tangent vector field ξ of M we shall express the Lie derivative of these basis vectors along ξ in terms of the same basis vectors.

Considering the orthonormal frame as orthogonal matrix, its derivative is a skew-symmetric matrix. Indeed, derivating the identity $\langle e_i , e_j \rangle = \delta_{ij}$ we have $\langle \partial_\xi e_i , e_j \rangle + \langle e_i , \partial_\xi e_j \rangle = 0$. Since the entries of this matrix depend linearly on ξ the directional derivatives have the form

$$\begin{array}{rcrcrcr} \partial_\xi e_1 &=& 0\, e_1 &+& a(\xi) e_2 &+& b(\xi) n\,, \\ \partial_\xi e_2 &=& -\ a(\xi) e_1 &+& 0\, e_2 &+& c(\xi) n\,, \\ \partial_\xi n &=& -\ b(\xi) e_1 &-& c(\xi) e_2 &+& 0\, n\,. \end{array} \tag{3}$$

The 1-forms a, b, c (which we shall determine) involve the second quadratic form h and the connection 1-form α. To express this involvement, denote $h_i(\xi) := h(\xi, e_i)$ and $h_{ji} = h(e_j, e_i) = h_i(e_j)$.

Lemma. *The required* 1*-forms are* $a = \alpha$, $b = h_1$, $c = h_2$, *that is,*

$$\begin{array}{rcrcrcr} \partial_\xi e_1 &=& 0\, e_1 &+& \alpha(\xi) e_2 &+& h_1(\xi) n\,, \\ \partial_\xi e_2 &=& -\ \alpha(\xi) e_1 &+& 0\, e_2 &+& h_2(\xi) n\,, \\ \partial_\xi n &=& -\ h_1(\xi) e_1 &-& h_2(\xi) e_2 &+& 0\, n\,. \end{array} \tag{4}$$

Proof. Observing that $\partial_\xi n = \Gamma_*(\xi)$ and $h(\xi,\eta) = -\langle\, \Gamma_*(\xi)\,,\, \eta\,\rangle$, we have $h_1(\xi) = -\langle\, \Gamma_*(\xi)\,,\, e_1\,\rangle$, $h_2(\xi) = -\langle\, \Gamma_*(\xi)\,,\, e_2\,\rangle$, that is,

$$\partial_\xi n = -h_1(\xi)e_1 - h_2(\xi)e_2\,.$$

Writing $a = a_1u_1 + a_2u_2$ it remains to prove $a_1 = \alpha_1$, $a_2 = \alpha_2$. Using that $b = h_1$ and $c = h_2$, set $\xi = e_2$ in the first equation of (3) and $\xi = e_1$ in the second one. Then subtract the first equation from the second to get

$$[e_1, e_2] = -a_1e_1 - a_2e_2 + (h_{21} - h_{12})n = -a_1e_1 - a_2e_2. \tag{5}$$

This commutator is completely determined by the vector fields e_1, e_2 on M, being independent of the embedding of M into $\mathbb{R}^3$. Relation (8) of p. 241 implies that the exterior derivative of any 1-form $u \in \Omega^1$ applied to any pair of vector fields ξ, η satisfies the following equality

$$du(\xi,\eta) = \partial_\xi u(\eta) - \partial_\eta u(\xi) - u([\xi,\eta])\,. \tag{6}$$

Hence, using (6) for $u = u_1$ and, then, applying (5), we get the equalities

$$\alpha_1 = du_1(e_1,e_2) = \partial_{e_1}u_1(e_2) - \partial_{e_2}u_1(e_1) - u_1([e_1,e_2]) = -u_1([e_1,e_2]) = a_1.$$

In the same way we get $\alpha_2 = a_2$, proving the Lemma. $\square$

Proof of Gauss' Theorem (ii). We shall apply both sides of the identity $\partial_{[e_1,e_2]} = \partial_{e_1}\partial_{e_2} - \partial_{e_2}\partial_{e_1}$ to e_1 and take the coefficient of e_2:

$$\left\langle\, \partial_{[e_1,e_2]}e_1\,,\, e_2\,\right\rangle = \left\langle\, \partial_{e_1}\partial_{e_2}e_1 - \partial_{e_2}\partial_{e_1}e_1\,,\, e_2\,\right\rangle.$$

For the left-hand side, set $\xi = [e_1, e_2]$ in the first equation of (4) to get

$$\left\langle \partial_{[e_1,e_2]}e_1\,,\, e_2 \right\rangle = \alpha([e_1,e_2]).$$

To compute the right-hand side, set $\xi = e_2$ in the first equation of (4), then set $\xi = e_1$ in the second equation to get

$$\left\langle\, \partial_{e_1}\partial_{e_2}e_1\,,\, e_2\,\right\rangle = \left\langle\, \partial_{e_1}\big(\,\alpha(e_2)\,e_2 + h_1(e_2)\,n\,\big)\,,\, e_2\,\right\rangle = \partial_{e_1}\alpha(e_2) - h_{12}h_{21}\,.$$

Similarly we obtain $\langle\, \partial_{e_2}\partial_{e_1}e_1\,,\, e_2\,\rangle = \partial_{e_2}\alpha(e_1) - h_{11}h_{22}$. We have proved that

$$\alpha([e_1,e_2]) = \partial_{e_1}\alpha(e_2) - \partial_{e_2}\alpha(e_1) + h_{11}h_{22} - h_{12}h_{21}\,,$$

which (setting $u = \alpha$, $\xi = e_1$, $\eta = e_2$ in relation (6)) is equivalent to

$$-d\alpha(e_1, e_2) = \det ||h_{ij}||\,, \quad \text{that is,} \quad K = \det ||h_{ij}|| = k_1 k_2\,. \qquad \square$$

Explicit formula for K. Since u_1, u_2 is a base, the differentials $d\alpha_1$, $d\alpha_2$ can be written as $d\alpha_1 = \alpha_{11}u_1 + \alpha_{12}u_2$ and $d\alpha_2 = \alpha_{21}u_1 + \alpha_{22}u_2$ for some functions α_{ij} (namely $\alpha_{ij} = \partial_{e_j}\alpha_i$).

Exercise. Compute $d\alpha$ to prove that the Gaussian curvature is given by

$$K = \alpha_{12} - \alpha_{21} - (\alpha_1^2 + \alpha_2^2)\,. \tag{7}$$

Problem. Compute the curvature of the conformal metric $g = \rho^2(dx^2 + dy^2)$.

Answer. $K = -\dfrac{\Delta \log \rho}{\rho^2}$, where Δ is the Laplacian operator $\dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2}$.

Hint. Use (7) and observe that $\frac{\partial}{\partial x}\left(\frac{\rho_x}{\rho}\right) = \frac{\rho\rho_{xx} - \rho_x^2}{\rho^2}$ and $\frac{\partial}{\partial y}\left(\frac{\rho_y}{\rho}\right) = \frac{\rho\rho_{yy} - \rho_y^2}{\rho^2}$.

9.3 The parallel transport

To define the curvature of a higher dimensional Riemannian manifold M^n we shall start with the definition of "parallel transport", as we did for surfaces ($n = 2$) in Section 8.4.

For $n > 2$ the difficulty comes from the fact that in this case the transport of the tangent vector of a geodesic curve along that geodesic, does not define the transport of the other tangent vectors. Of course, the parallel transport should be an isometric linear map from the tangent space of the Riemannian manifold at the initial point to the tangent space at the final point.

Thus, the angle between the transported vector and the tangent vector orienting the geodesic line at the final point is known: It is the same as for the initial vector at the initial point.

In the case of the surfaces, $n = 2$, this condition determines the transported vector, but already for $n = 3$ the vectors forming a prescribed angle with the direction of the geodesic line form a cone, and we have to choose from it the correct transported vector.

To do this, it would suffice to know the 2-dimensional plane generated by the vector orienting the geodesic line and the transported vector, in the tangent space of M^n at the final point. We shall now define geometrically this special plane – Fig. 9.4.

On the tangent space of M^n at the initial point x, T_xM^n, we consider the 2-dimensional subspace V generated by a vector ξ and the tangent vector v of the geodesic γ along which we shall transport the vector ξ (Fig. 9.4).

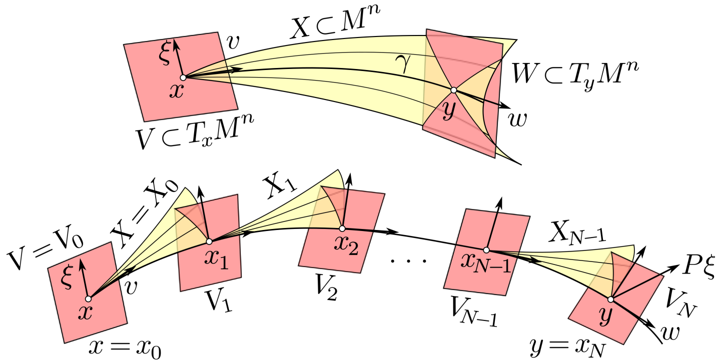


Figure 9.4: Parallel transport along the geodesic γ from the point $x = x_0$ to the point $y = x_N$ on a higher-dimensional Riemannian manifold.

The geodesics of M^n that start at the point x in directions lying in the plane V form a smooth 2-dimensional surface (in a neighbourhood of the initial point x) that we denote by $X \subset M^n$. This surface contains the geodesic γ.

Let y be a point of γ sufficiently close to x. The tangent space of the smooth surface X at y is a 2-plane $W \subset T_yM^n$ containing the tangent vector of γ at y. To define parallel transport of vectors of the initial plane $V \subset T_xM^n$ to T_yM^n, we shall use a 2-plane of T_yM^n *different* from W

Namely, we subdivide the segment $[x, y]$ of γ into N small pieces (say, of equal lengths) by points x_k ($0 \leqslant k \leqslant N$), denoting x by x_0 and y by x_N, and then we construct inductively a sequence of 2-planes $V_k \subset T_{x_k}M^n$, starting from the 2-plane $V_0 = V$ at x, by the following method:

The geodesics starting at x_k in the directions tangent to the plane V_k that we want to define, form a smooth surface X_k containing the segment $[x_k, x_{k+1}]$ of the geodesic γ (see Fig. 9.4). We define the 2-plane V_k as the tangent plane to the surface X_{k-1} at the point x_k.

The resulting plane V_N at y depends on N, but there exist a limit plane

$$\widetilde{V} = \lim_{N\to\infty} V_N \subset T_yM^n.$$

Definition. The parallel transport of the vector ξ of the plane V at x along the geodesic γ (to the point y) is the vector $P\xi$ of $\widetilde{V}$, whose (oriented) angle with the tangent vector of γ at the final point y is equal to the angle of ξ with the direction of γ at the initial point x (and whose length is equal to the length of the initial vector).

The linear map $P : T_xM^n \to T_yM^n$ is orthogonal and depends smoothly on y.

Problem. Transport parallelly a vector along the meridian of the standard sphere $\mathbb{S}^n$ (from one pole to the other).

Problem. Transport parallelly a vector along the geodesic line $b = 0$ of the n-dimensional Lobachevsky space $\{a, b : a > 0, b \in \mathbb{R}^{n-1}\}$, equipped with the metric of the Poincaré model:

$$(ds)^2 = \frac{(da)^2 + \sum_{k=1}^{n-1}(db_k)^2}{a^2}.$$

Problem. Transport parallelly a vector along the geodesic line of the complex projective plane $M^4 = \mathbb{C}\mathrm{P}^2$, equipped with the following (standard) metric: Understanding the space $\mathbb{C}\mathrm{P}^n$ as consisting of the circles $\mathbb{S}^1$ (along which the unit sphere $\mathbb{S}^{2n+1}$ of the complex Hermitian vector-space $\mathbb{C}^{n+1}$ intersects the complex 1-dimensional subspaces), we define the distance ds between two (infinitely close) points of $\mathbb{C}\mathrm{P}^n$ as the distance between the two (infinitely close) corresponding circles in the sphere.

Problem. Calculate the area of $\mathbb{C}\mathrm{P}^1$ and the 4-volume of $\mathbb{C}\mathrm{P}^2$ with these Riemannian metrics.

Remark. It is also useful to write the explicit formula for the quadratic form $(ds)^2$ in terms of the affine coordinates z_k, $1 \le k \le n$, in $\mathbb{C}\mathrm{P}^n$.

Hint. If the tangent vector $\zeta \in T_x\mathbb{C}\mathrm{P}^n$ is the natural projection of the tangent vector $\xi \in \mathbb{C}^{n+1}$, then

$$(ds)^2(\zeta) = \frac{\langle \xi, \xi\rangle\langle z, z\rangle - \langle \xi, z\rangle\langle z, \xi\rangle}{\langle z, z\rangle^2},$$

where $\langle\cdot,\cdot\rangle$ is the complex Hermitian scalar product.

Note that this metric is not invariant with respect to the complex projective transformations of the complex projective space. It is, however, invariant under those of them which are induced by the unitary transformations of the Hermitian space $\mathbb{C}^{n+1}$.

Returning to the general theory, we define the parallel transport along any smooth curve in a Riemannian manifold M^n exactly in the same way as in the 2-dimensional case (see Fig. 8.9, p. 283): Subdivide the segment $[x, y]$ of the curve into N small segments, replace each of these small segments by a (short) geodesic line and transport the original vector along the resulting broken geodesic line. The parallel transport along the original curve is the limit of the parallel transport along these broken geodesic lines for $N \to \infty$.

Problem. Transport a vector along the line $a = 1$, $b_2 = \ldots = b_n = 0$ of the Poincaré model of the n-dimensional Lobachevsky space (see p. 327).

We have now two definitions of the parallel transport along the curves on the Riemannian manifolds: Either to embed the manifold with its metric to the Euclidean space with its metric, and to project orthogonally the affine tangent space at the initial point onto the affine tangent space at the next (infinitesimally close) point of the curve, or to transport along the broken geodesic lines, as it is explained above.

We have not proved that both ways lead to the same result, but it is true. Hence, the method of the affine projections provides the intrinsic transport, independently of the choice of the isometric embedding to the Euclidean space, on which it is based and on which the result might depend.

Radon's Parallel Transport. The mathematician Radon invented, about 1918, a third way to define the same parallel transport, which, physically, is the most natural one. As far as we know, it was not published by Radon, but quoted by F. Klein in his book *Vorlesungen über höhere Geometrie*, Berlin, Grunlagen Math-Wissenschaften, Einzendarstellung 1926.

Radon started with a generalised "Foucault pendulum", which is an oscillatory system at a point x of M^n, whose potential energy U is defined as

$$U(q) = \frac{(\text{distance from } q \text{ to } x)^2}{2},$$

and the kinetic energy of the point moving along the Riemannian manifold is the usual $(|\dot{q}|^2)/2$. It suffices even to understand $\dot{q}$ as the tangent vector at x, and to take the quadratic potential energy U.

The motion of this "spherical pendulum" is periodic, following a line ℓ of $T_x M^n$ if the initial velocity is zero.

To define the parallel transport of a vector of the line ℓ along a smooth curve γ, from the point x to the point y, Radon suggested the "adiabatic process" of a slow variation of the spherical pendulum system described above. Namely, one moves *slowly* the point x along the curve γ (say, as $x(t) = \gamma(\varepsilon t)$ for the parametrised curve γ).

The resulting non autonomous (linear) system defines, at every moment t, some spherical pendulum at some point $x(t)$, and the motion in the non autonomous system is, at this moment, very close to the periodic motion $\xi \cos t$ for some vector $\xi \in T_{x(t)}M$.

Rigorously speaking, one should fix the final point $x(t) = y$ and consider the limit of the vector $\xi \in T_{x(t)}M$ for $\varepsilon \to 0$ (provided by the "adiabatic averaging" theory of the slowly varying dynamical systems). The obtained vector ξ in T_yM provides the direction of the (parallelly) transported vector. Parallel transport is in this sense *the adiabatic transport*.

9.4 The notion of curvature

Having defined the parallel transport, we turn to the notion of curvature. As in the 2-dimensional case, we start from the "transport along an infinitesimal parallelogram" defined by two tangent vectors $\xi, \eta \in T_xM^n$ (Fig. 9.5).

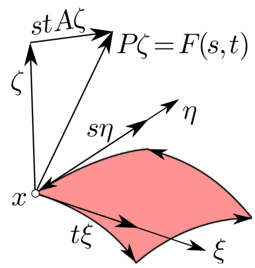


Figure 9.5: The parallel transport of the tangent vector ζ (to $P\zeta = F(s,t)$) along the infinitesimal parallelogram with sides $t\xi$ and $s\eta$, on a Riemannian manifold.

Let ζ be a third vector of the tangent space T_xM^n. We will transport it along the boundary of the 2-chain defined by the two small vectors $t\xi$, $s\eta$.

The transported vector $P\zeta$ belongs to the same space T_xM^n, as the vector ζ, and depends on the small parameters s and t: $P\zeta = F(s,t)$.

For $s = 0$ and for $t = 0$ the transported vector $P\zeta$ is ζ, and hence the Taylor series of F starts with the term ζ, followed by a bilinear term:

$$F(s,t) = \zeta + stA\zeta + O(|s|^3 + |t|^3).$$

Since the parallel transport preserves the metric, it follows that the linear operator A is antisymmetric (since it belongs to the tangent space of the special orthogonal group $\mathrm{SO}(n)$, considered on p. 255).

This linear operator depends anti-symmetrically on the two vectors ξ and η of the tangent space. These dependences are linear, and we can consider $A = A(\xi, \eta)$ as being the value, at the pair (ξ, η) of a generalised differential 2-form Ω on M^n (whose values belong to the Lie algebra $\mathfrak{o}(n)$ of the orthogonal group of the fibre T_xM^n).

To write explicitly the dependence of the linear operator A on the tangent vectors ξ and η, one considers the three-linear vectorial function with value

$$\Omega(\xi, \eta)\zeta,$$

as a tensor map

$$\Omega : V^n \times V^n \times V^n \to V^n,$$

where $V^n = T_x M^n$.

To return to the notations with indices and matrices, one fixes one more vector ν in the tangent space V^n and defines the number

$$R(\xi, \eta, \zeta, \nu) = \langle \Omega(\xi, \eta)\zeta, \nu \rangle \in \mathbb{R}$$

as the scalar product of the vector $\Omega(\xi, \eta)\zeta$ with the vector ν.

Riemann Tensor. The 4-linear form

$$R : V^n \times V^n \times V^n \times V^n \to \mathbb{R}$$

is called *the Riemann tensor* of the Riemannian manifold M^n at x.

Fixing a Cartesian orthonormal basis $(e_1, \dots, e_n)$ in V^n, we obtain the "components" of the Riemann tensor,

$$R_{i,j,k,l} = R(e_i, e_j, e_k, e_l).$$

All geometric properties of the parallel transport are reflected in some relations between these n^4 functions (see the end of this chapter for more relations).

Example. The anti-symmetry in ξ and η: $R_{i,j,k,l} = -R_{j,i,k,l}$.

Moreover, $R_{i,j,k,l} = -R_{i,j,l,k}$, since $\Omega(\xi, \eta)$ belongs to the Lie algebra $\mathfrak{o}(n)$.

In the case $n = 2$, we have seen some other properties (say, the closeness of the curvature form) which are particular cases of similar properties of the Riemann tensor for the general case. The proofs are not difficult, provided that you are carefully controlling the notations, the indices and the signs. Einstein had some difficulties dealing with these "Bianchi identities", in his works on the general relativity theory, where $n = 4$, but Hilbert helped him to write correctly this variational principle. The discovery (made later) of the equivalence of Hilbert's principle with that of Einstein is based on these identities between the coefficients $R_{i,j,k,l}$ of the Riemann tensor.

The story of the discovery of all these objects starts from the attempts of Riemann to classify the Riemannian metrics locally, up to isometries. Say, for surfaces, the spheres of different radii are not locally isometric to each other, and the Lobachevsky plane also differs from them; while the plane, the cones and the cylinders are all locally isometric. In these cases the Gaussian curvature is the invariant distinguishing non isometric surfaces.

The problem of Riemann was to extend this invariant to manifolds of higher dimensions. So he tried to answer the question: Which properties of Riemannian manifolds do depend on the embedding in Euclidean spaces (like the principal curvatures of surfaces in $\mathbb{R}^3$) or depend on the coordinate systems, and which of them are true inner properties?

These questions of Riemann (which originated from his attempts to understand the geometric structures of physical fields) are the direct ancestors of the relativity theory of Poincaré (who formulated it in 1895) and A. Einstein (who published in 1905 his translation of the rather philosophic and geometric paper of Poincaré in the language of formulas and indices).

To realise his program, Riemann started with the following calculations. The metric on M^n is locally defined by the $n(n+1)/2$ components of the quadratic form $(ds)^2$, which are functions, denoted by $g_{i,j}(x)$, of the n coordinates $(x_1, \ldots, x_n)$ of the point $x \in M^n$.

This description is coordinate-dependent: Choosing a different coordinate system $(\widetilde{x}_k)$ (related to the system (x_k) by a diffeomorphism defined by n functions $x_k = f_k(\widetilde{x})$ of n variables), we get a new description $\widetilde{g}_{i,j}(\widetilde{x})$ of the old quadratic form defining the metric $(ds)^2$. Hence the problem of Riemann was to find which terms of the Taylor series of the coefficients $g_{i,j}$ at the given point (say, at the origin $x = 0$) may be transformed by a suitable choice of the coordinate system, and which can not.

For the term of degree 0 the solution is given by the normal forms of the theory of quadratic forms: All metrics are reducible to the form

$$(ds)^2 = \sum_{k=1}^{n} (dx_k)^2,$$

corresponding to the unit matrix $(g_{i,j}(x)) = 1 + O(|x|)$, perturbed by the first order terms.

Thus one tries to reduce these first order terms by non-linear changes of variables. Doing these calculations Riemann discovered his curvature tensor R, which is an invariant of the metric and is geometrically defined: Although the coefficients $R_{i,j,k,l}$ may be changed by different choices of the coordinate systems, the geometric operator $\Omega(\xi,\eta): T_xM^n \to T_xM^n$ remains the same, and the differential 2-form Ω with values in $\mathrm{O}(n)$ is independent of any coordinate system. The dependence of the coefficients $R_{i,j,k,l}$ on the coordinate system is a manifestation of the dependence of the elements of a matrix on the chosen basis.

In this way Riemann discovered the invariant of minimal order of the metric at a point (extending to higher dimensions the Gaussian curvature).

9.5 Riemann sectional curvature

Definition. The *Riemann sectional curvature* (of a Riemannian manifold at a point x) in the direction of a 2-dimensional plane $V \subset T_xM^n$ is the real number

$$K(\xi,\eta) = \langle \Omega(\xi,\eta)\xi, \eta \rangle,$$

where ξ and η are two orthogonal unit vectors of the plane V.

This number depends on the plane V, but is independent of the orthonormal basis (ξ,η) (of V) chosen to calculate it. The value of the sectional curvature can be written in the form $K(\xi,\eta) = R_{i,j,i,j}$, where $\xi = e_i$ and $\eta = e_j$ are the basic vectors of the Cartesian orthonormal coordinate system at T_xM^n.

2nd definition. The *sectional curvature* $K(\xi,\eta)$ is equal to the Gaussian curvature of the 2-dimensional submanifold formed by the geodesics of M^n which start at x in all the directions belonging to the plane generated by the vectors ξ and η of T_xM^n.

Problem. Prove that both definitions of sectional curvature are equivalent.

Problem. Calculate the sectional curvatures of the sphere $\mathbb{S}^n$ of radius r in Euclidean space $\mathbb{R}^{n+1}$.

Answer. All sectional curvatures of such a sphere are equal to $1/r^2$.

Problem. Calculate the sectional curvatures of the Lobachevsky n-space (whose metric is defined on p. 327 for the Poincaré model).

ANSWER. All sectional curvatures of Lobachevsky n-space are equal to -1 (but if we multiply the distances ds by a constant c, then the sectional curvature will become $-1/c^2$).

PROBLEM. Calculate the sectional curvatures of the 4-dimensional manifold $M^4 = \mathbb{C}\mathrm{P}^2$ (whose metric is discussed on p. 327).

Hint. It is non negative. The exact value depends on the position of the plane $V \subset T_x M^4$ with respect to the complex subspaces of the complex vector-space $T_x(\mathbb{C}\mathrm{P}^2)$. We give the solution in the following section.

9.5.1 Sectional curvatures of $\mathbb{C}\mathrm{P}^2$

To measure the position of a plane $V \subset T_x M^4$ with respect to the complex subspaces, we use the

Definition. The *anomaly* α of a real 2-plane V in the complex Hermitian space $\mathbb{C}^2$ is the angle defined by the following construction. Take any non-zero vector v in V. Construct the non-zero orthogonal vector w in V and compare it with the vector iv. The anomaly α is the angle between the vectors w and iv:

$$\cos\alpha = \frac{\mathrm{Re}\,\langle w, iv\rangle}{|w||v|}.$$

The anomaly angle α is independent of the choice of v in V, namely, the real number $\cos\alpha$ is the same for any v, and hence it determines α up to the sign. We leave to the reader the easy verification of this independence.

Example. The anomaly of a complex subspace $V \approx \mathbb{C}^1$ equals zero, since, in this case, iv belongs to V, making $w = \pm iv$.

Example. The anomaly of the subspace of real directions $V = \mathbb{R}\mathbf{e} + \mathbb{R}\mathbf{f}$ of the complex plane $\mathbb{C}^2$ with Hermitian basis $(\mathbf{e}, \mathbf{f})$, is equal to $\pm\pi/2$. In this case, for $v = \mathbf{e}$ we find $w = \mathbf{f}$, which is orthogonal to $iv = i\mathbf{e}$.

Example. The anomaly of the real plane $V = \{z\mathbf{e} + w\mathbf{f} : w = (\tan\varphi)z\}$ is $\alpha = 2\varphi$. Indeed, for $v = e^{i\beta}\mathbf{e} + (\tan\varphi)e^{-i\beta}\mathbf{f}$, we get $w = ie^{i\beta}\mathbf{e} - i(\tan\varphi)e^{-i\beta}\mathbf{f}$, so

$$\langle w, iv\rangle = 1 - \tan^2\varphi, \qquad |v|^2 = |w|^2 = 1 + \tan^2\varphi,$$

and hence

$$\cos\alpha = (1 - \tan^2\varphi)/(1 + \tan^2\varphi) = \cos^2\varphi - \sin^2\varphi = \cos 2\varphi.$$

In fact, any plane of anomaly $\pm 2\varphi$ is provided by this example, for a suitable choice of the Hermitian basis $\{e, f\}$ in $\mathbb{C}^2$.

Theorem 1. *The sectional curvature of* $\mathbb{C}\mathrm{P}^2$ *in the direction of a plane* V *with anomaly* α *equals* $K(V) = 3\cos^2\alpha + 1$.

Example. The sectional curvature along any complex line is equal to 4, since $\alpha = 0$. It is conformal to the distance $\pi/2$ between the opposite points of the sphere $\mathbb{C}\mathrm{P}^1$, represented by two Hermitian orthogonal lines $\mathbb{C} \subset \mathbb{C}^2$. Hence the radius of this sphere equals $1/2$. The sectional curvature along a subspace of real directions is equal to 1, since $\alpha = \pm\pi/2$. Thus "the radius of curvature along the real directions" is two times lager than "the radius of curvature along the complex directions".

The sectional curvature at a point, in the direction of a plane V, is equal to the curvature of the real surface formed by the geodesic lines starting at that point along the section V.

To calculate the curvature of a real surface, suppose that its metric is defined by the formula

$$ds^2 = (dz)^2 + (dw)^2 + A(dz)^2 + 2B dz dw + C(dw)^2$$

in some real coordinates (z, w) centred at the point where the curvature is calculated. The formula represents a perturbation of the Euclidean metric $(dz)^2 + (dw)^2$, and it is easy to choose some coordinates for which the metric will have the above expression, with $A(0) = B(0) = C(0) = 0$ (it is the theorem on the normal form of the Euclidean structure).

Next, making a small diffeomorphism preserving the origin, we can kill the first order terms of the coefficients A, B, C: It suffices to put $z = \widetilde{z} + a_2(\widetilde{z}, \widetilde{w})$, $w = \widetilde{w} + b_2(\widetilde{z}, \widetilde{w})$, where a_2 and b_2 are homogeneous polynomials of degree 2 (their 6 coefficients suffice to kill the 6 coefficients of the linear functions of two variables).

Hence, to calculate the Gaussian curvature of an arbitrary metric, it suffices to consider the case where the perturbing coefficients A, B, C are quadratic homogeneous polynomials:

$$A = A_{2,0}z^2 + 2A_{1,1}zw + A_{0,2}w^2,$$

$$B = B_{2,0}z^2 + 2B_{1,1}zw + B_{0,2}w^2,$$

$$C = C_{2,0}z^2 + 2C_{1,1}zw + C_{0,2}w^2.$$

Lemma 1. *In this case the Gaussian curvature of the surface, at the origin, is*

$$K = 2B_{1,1} - A_{0,2} - C_{2,0}.$$

Proof of Lemma 1. Use the diffeomorphism $(z = \widetilde{z} + a_3(\widetilde{z}, \widetilde{w}),\ w = \widetilde{w} + b_3(\widetilde{z}, \widetilde{w}))$, where

$$a_3 = D\widetilde{z}^3 + E\widetilde{z}^2\widetilde{w} + F\widetilde{z}\widetilde{w}^2 + G\widetilde{w}^3,$$
$$b_3 = P\widetilde{z}^3 + Q\widetilde{z}^2\widetilde{w} + R\widetilde{z}\widetilde{w}^2 + S\widetilde{w}^3.$$

The 8 coefficients $(D, \ldots, S)$ allow us to kill 8 of the 9 coefficients $(A_{2,0}, \ldots, C_{0,2})$, and the remaining coefficient defines the curvature K.

In fact this reasoning was the method used by Riemann to define his curvatures: They were discovered as the coefficients of the smallest degree, in the Taylor series of the Riemannian metric, which can't be annihilated by a choice of the coordinate system

We get explicitly the expression of the metric in the new coordinates

$$ds^2 = (d\widetilde{z})^2 + (d\widetilde{w})^2 + \widetilde{A}(d\widetilde{z})^2 + 2\widetilde{B}d\widetilde{z}d\widetilde{w} + \widetilde{C}(d\widetilde{w})^2,$$

where the Taylor series of the perturbing coefficients $\widetilde{A}$, $\widetilde{B}$, $\widetilde{C}$ start from the quadratic terms

$$\widetilde{A_2} = \widetilde{A_{2,0}}\widetilde{z}^2 + 2\widetilde{A_{1,1}}\widetilde{z}\widetilde{w} + \widetilde{A_{0,2}}\widetilde{w}^2,$$

and similarly for B_2 and C_2.

To calculate these coefficients we use the differentials

$$dz = (1 + 3D\widetilde{z}^2 + 2E\widetilde{z}\widetilde{w} + F\widetilde{w}^2)d\widetilde{z} + (E\widetilde{z}^2 + 2F\widetilde{z}\widetilde{w} + 3G\widetilde{w}^2)d\widetilde{w}$$

$$dz = (3P\widetilde{z}^2 + 2Q\widetilde{z}\widetilde{w} + R\widetilde{w}^2)d\widetilde{z} + (1 + Q\widetilde{z}^2 + 2R\widetilde{z}\widetilde{w} + 3S\widetilde{w}^2)d\widetilde{w},$$

which provide the 9 values of the new coefficients

$$\widetilde{A_{2,0}} = A_{2,0} + 6D, \qquad 2\widetilde{A_{1,1}} = 2A_{1,1} + 4E, \qquad \widetilde{A_{0,2}} = A_{0,2} + 2F,$$

$$2\widetilde{B_{2,0}} = 2B_{2,0} + 2E + 6P, \quad 4\widetilde{B_{1,1}} = 4B_{1,1} + 4F + 4Q, \quad 2\widetilde{B_{0,2}} = B_{0,2} + 2G + 2R;$$

$$\widetilde{C_{2,0}} = C_{2,0} + 2Q, \qquad 2\widetilde{C_{1,1}} = 2C_{1,1} + 4R, \qquad \widetilde{C_{0,2}} = C_{0,2} + 6S.$$

One choose $D = -A_{2,0}/6$, $E = -A_{1,1}/2$, $F = -A_{0,2}/2$, $Q = -C_{2,0}/2$, $R = -C_{1,1}/2$, $S = -C_{0,2}/6$, $P = -(B_{2,0} + E)/3$, $G = -(B_{0,2} + R)/3$, to kill all the new coefficients except

$$4\widetilde{B_{1,1}} = 4B_{1,1} + 4F + 4Q = 4B_{1,1} - 2A_{2,0} - 2C_{2,0}.$$

Example. Consider the surface in Euclidean 3-space defined, in the Cartesian orthonormal coordinates (z, w, t), by the equation

$$t = p\frac{z^2}{2} + q\frac{w^2}{2}.$$

Its principal curvatures at the origin are p and q. Hence its Gaussian curvature is $K = pq$. Using the differential $dt = pzdz + qwdw$ and applying the above calculations, we get

$$\begin{aligned} ds^2 &= (dz)^2 + (dw)^2 + (dt)^2 \\ &= (dz)^2 + (dw)^2 + p^2z^2(dz)^2 + 2pqzw(dz)(dw) + q^2w^2(dw)^2, \end{aligned}$$

that is, $A_{0,2} = 0$, $C_{2,0} = 0$, and $4B_{1,1} = 2pq$. Hence $4\widetilde{B_{1,1}} = 2pq$.

It means that every surface is isometric to the surface of our example, with coefficients $pq = 2B_{1,1} - A_{2,0} - C_{2,0}$, modulo higher order terms in the Taylor series (which will not influence the Gaussian curvature at the origin).

Thus the Gaussian curvature of an arbitrary surface, at the origin, is equal to the Gaussian curvature $K = pq$ of the special surface of the example. Lemma 1 is proved. $\square$

To apply the formula of Lemma 1 to the surface of $\mathbb{C}\mathrm{P}^2$ formed by the geodesic lines starting at one point and going along the directions of a real 2-plane V, we shall use a general result of variational calculus, formulated below for the case of the Riemannian metric (for which we shall use it).

Consider an m-dimensional Riemannian manifold M, equipped with a one-parameter family $\{g^t\}$ of isometric diffeomorphisms onto itself (a “flow”), whose orbits $\{g^t x : t \in \mathbb{R}\}$ are geodesic lines of M.

The space of the orbits $\widehat{M}$ (defined at least locally) inherits a Riemannian metric for which the infinitesimal distance between two neighbouring points of $\widehat{M}$ is the Riemannian distance between the neighbouring orbits.

Remark. Notice that to measure the distance between orbits it is enough to take the local distance. In general, the orbits could fill the manifold densely, so that the global distance could be zero even for non-coinciding orbits (and globally the space of orbits is not a manifold).

When we discuss the “local” orbit space we mean that every arc of the intersection of an orbit with the given neighbourhood is considered as a separate point of $\widehat{M}$ (even if two such “points” correspond globally to the same orbit).

This kind of pathology does not happen in the considered example of $\mathbb{C}\mathrm{P}^n$, but the idea of the quotient Riemannian metric is more general.

To measure the distance between two orbits, we take a point on the first orbit and look for the closest point on the second one. This does not depend on the choice of the point of the first orbit, since the group acts by isometries.

Consider now the natural projection $\pi : M \to \widehat{M}$, sending any point x of M to its orbit $\{g^t x : t \in \mathbb{R}\}$.

Example. Let $M = \mathbb{S}^{2n+1}$ be the sphere of unit vectors in Hermitian space $\mathbb{C}^{n+1}$. Taking the natural Hopf fibration $\pi : \mathbb{S}^{2n+1} \to \mathbb{C}\mathrm{P}^n$, the diffeomorphism g^t is defined by the multiplication by e^{it}, which is an isometry for any $t \in \mathbb{R}$.

The general result we shall use is the

Lemma 2. *If a geodesic line of M is orthogonal to the orbit of the flow at one of its points, then it is orthogonal to the orbits of the flow at each of its points, and the projection $\pi : M \to \widehat{M}$ sends it to a geodesic line of $\widehat{M}$.*

Proof. It suffices to write the Lagrange equations $\frac{d}{dt}(\partial L/\partial \dot{q}) = \partial L/\partial q$ for the geodesic lines on M and $\frac{d}{dt}(\partial \widehat{L}/\partial \dot{\widehat{q}}) = \partial \widehat{L}/\partial q$ for the geodesic lines on $\widehat{M}$. We shall use that the differential of the restriction of a function coincides with the restriction of the differential of the unrestricted function. We have here that the kinetic energy function $\widehat{L}$ for the quotient manifold $\widehat{M}$ is the restriction of the kinetic energy function L for the initial Riemannian manifold M. Namely, to get the value of the quadratic form $\widehat{L}$ on a tangent vector $\dot{\widehat{q}} \in T_{\widehat{q}}\widehat{M}$ we have to consider any point q of the (geodesic) orbit $\pi^{-1}\widehat{q} \subset M$ and to lift the tangent vector $\dot{\widehat{q}}$ to a vector $\dot{q} = \xi$ in the orthogonal subspace $W \subset T_q M$ to that orbit at q.

By the definition of the quotient metric on M, we have $\widehat{L}(\dot{\widehat{q}}) = L(\xi)$; and the orthogonality condition along the fibre is $p(\xi) = 0$ for $p(\xi) = \partial L/\partial \dot{q}_{|\dot{q}=\xi}$.

Hence the Lagrange equation on M provides $dp/dt = 0$ for the motion along the geodesic orthogonal to the fibre, while $p(\xi) = 0$ for the initial condition at the starting point of this geodesic. Thus $p(t) \equiv 0$, and hence the geodesic orthogonal to the orbit at the initial point is orthogonal to the orbits at each of its points. Lemma 2 is proved. $\square$

Remark. A reader unsatisfied with this simple reasoning can write explicitly the Lagrange equations, choosing some local coordinates $x = (x_1, \ldots, x_{m-1})$ in $\widehat{M}$, and taking $(x, y) \in \mathbb{R}^m$ as the local coordinates of $g^y(q(x)) \in M$ for some transverse section $q : \widehat{M} \to M$.

Making no mistakes in the fighting with the indices, one gets the above conclusions from the kinetic energy formula

$$L(x, y, \dot{x}, \dot{y}) = \frac{1}{2}(A\dot{x}, \dot{x}) + (b\dot{x})\dot{y} + \frac{1}{2}c\dot{y}^2,$$

whose coefficients (the symmetric matrix A, the vector b and the number c) do not depend on y, since the symmetry $g^t(x, y) = (x, y+t)$ preserves L.

In these notations the orthogonality condition takes the form $b\dot{x} + c\dot{y} = 0$. Hence the Lagrange function in the quotient space is

$$\begin{aligned}\widehat{L}(x, \dot{x}) &= \frac{1}{2}(A\dot{x}, \dot{x}) + (b\dot{x})\left(-\frac{b\dot{x}}{c}\right) + \frac{c(b\dot{x}/c)^2}{2} \\ &= \frac{1}{2}(A\dot{x}, \dot{x}) - \frac{1}{2}\frac{(b\dot{x})^2}{c}.\end{aligned}$$

The Lagrange equations for the x-variables in M and in $\widehat{M}$ are equivalent, provided that the orthogonality condition $\partial L/\partial \dot{y} = 0$ holds, that is, $b\dot{x} + c\dot{y} = 0$. The orthogonality condition is preserved, since $\frac{d}{dt}(\partial L/\partial \dot{y}) - \partial L/\partial y \equiv 0$.

Thus, to get the quotient metric on the surface formed by the geodesics starting from a point $\eta \in \mathbb{C}\mathrm{P}^2$ along the directions of a 2-plane $V \subset T_\eta \mathbb{C}\mathrm{P}^2$, it suffices to lift those geodesics to some geodesics of the sphere $\mathbb{S}^5$, orthogonal to fibres of the Hopf fibration $\pi : \mathbb{S}^5 \to \mathbb{C}\mathrm{P}^2$. The resulting surface in $\mathbb{S}^5$ is provided by the following lemma.

Lemma 3. *The geodesics of the sphere* $\mathbb{S}^5 = \{z, w, t : |z|^2 + |w|^2 + |t|^2 = 1\}$, *which start at the point* $\xi = (0, 0, 1)$ *in the directions tangent to the plane* $(dw = (\tan\varphi)dz,\ dt = 0)$ *of* $T_\xi \mathbb{S}^5$, *form the surface* (parametrised by the two real variables τ and ϑ):

$$\{z = e^{i\vartheta}\cos\varphi \sin\tau, \quad w = e^{-i\vartheta}\sin\varphi\sin\tau, \quad t = \cos\tau\}.$$

To get the orthogonality of the geodesics to the fibre of the Hopf fibration, we have chosen a plane for which $dt = 0$. Any real 2-plane with anomaly $\alpha = 2\varphi$ and orthogonal to the fibre can be defined by the above equation, for a suitable choice of the Hermitian coordinates (z, w). Consequently, the example of Lemma 3 contains, up to the coordinate notations, all the possible cases.

Proof of Lemma 3. The geodesic of the unit sphere, which starts at the point $\xi \in \mathbb{S}^5$ with the velocity $\eta \in T_\xi \mathbb{S}^5$ $(|\eta| = 1)$, is the great circle

$$\{\xi\cos\tau + \eta\sin\tau, \ \ \tau \in \mathbb{R}\},$$

where $\langle \xi, \xi \rangle_{\mathbb{R}} = 1$, $\langle \eta, \eta \rangle_{\mathbb{R}} = 1$, $\langle \xi, \eta \rangle_{\mathbb{R}} = 0$ for the Euclidean scalar product $\langle \cdot, \cdot \rangle_{\mathbb{R}}$ in $\mathbb{R}^6$.

We choose the tangent vector η in the plane $dw = (\tan\varphi)dz$, $dt = 0$, for which $dz = e^{i\vartheta}$, $dw = e^{-i\vartheta}\tan\varphi$.

To obtain a vector of length 1, we normalise η multiplying it by $\cos\varphi$:

$$\eta_\vartheta = (z = e^{i\vartheta}\cos\varphi,\ w = e^{-ie}\sin\varphi,\ t = 0).$$

The great circles $\xi\cos\tau + \eta_\vartheta \sin\tau$ form the parametrised surface of Lemma 3, for the fixed anomaly $\alpha = 2\varphi$ of the initial tangent plane. □

Proof of Theorem 1. It remains to calculate the Gaussian curvature of the image of the surface of Lemma 3, under the Hopf projection from the standard unit 5-sphere $\mathbb{S}^5$ to the Riemannian manifold $\mathbb{C}\mathrm{P}^2$.

Denote by $r = r(\tau, \vartheta)$ the point under consideration of the surface, and by $\dot r$ the velocity of its motion along the surface, $\dot r = (\partial r/\partial\tau)\dot\tau + (\partial r/\partial\vartheta)\dot\vartheta$.

According to the formula of p. 327, the metric of the projected surface is

$$ds^2 = \langle \dot r, \dot r \rangle_{\mathbb{C}} - \langle \dot r, r \rangle_{\mathbb{C}} \langle r, \dot r \rangle_{\mathbb{C}},$$

for the Hermitian scalar product $\langle \cdot, \cdot \rangle_{\mathbb{C}}$ in $\mathbb{C}^3$. Here, we have used the identity $\langle r, r \rangle_{\mathbb{C}} = 1$, holding in our example.

To simplify the calculations, we will not parametrise the surface by the "polar" coordinates τ and ϑ, but by the "Cartesian" coordinates $a = \tau\cos\vartheta$, $b = \tau\sin\vartheta$, obtaining the parametrisation

$$r = \{\ z = (a + bi)\cos\varphi,\quad w = (a - bi)\sin\varphi,\quad t = 1 - \frac{a^2 + b^2}{2}\ \},$$

where we have neglected the cubic and higher order terms in a and b of the Taylor series of r. So, neglecting these small terms – which do not influence the curvature at 0 – we obtain

$$\dot r = \Big((\dot a + i\dot b)\cos\varphi,\ (\dot a - i\dot b)\sin\varphi,\ -a\dot a - b\dot b \Big),$$

$$\langle \dot r, \dot r \rangle = (\dot a^2 + \dot b^2)(\cos^2\varphi + \sin^2\varphi) + a^2\dot a^2 + b^2\dot b^2 + 2ab\dot a\dot b \qquad \text{and}$$

$$\begin{aligned} \langle \dot r, r \rangle &= ((a\dot a + b\dot b) + i(a\dot b - \dot a b))\cos^2\varphi + (\dot a a + b\dot b) + i(\dot a b - a\dot b))\sin^2\varphi - a\dot a - b\dot b \\ &= (a\dot a + b\dot b)(\cos^2\varphi + \sin^2\varphi - 1) + i(a\dot b - \dot a b)(\cos^2\varphi - \sin^2\varphi) \\ &= i(a\dot b - \dot a b)\cos 2\varphi. \end{aligned}$$

Thus, $\langle \dot{r}, r\rangle\langle r, \dot{r}\rangle = |\langle \dot{r}, r\rangle|^2 = (a^2\dot{b}^2 + b^2\dot{a}^2 - 2ab\dot{a}\dot{b})\cos^2 2\varphi$. Consequently, the metric of the surface formed by the geodesics in $\mathbb{C}\mathrm{P}^2$ is, modulo $O(|a|^2 + |b|^2)$:

$$\begin{aligned} ds^2 = (da)^2 + (db)^2 + a^2(da)^2 + 2ab(da)(db) + b^2(db)^2 \\ - (b^2(da)^2 + a^2(db)^2 - 2ab(da)(db))\cos^2 2\varphi. \end{aligned}$$

Now, to apply Lemma 1, we calculate the coefficients A, $2B$, C of the perturbing quadratic form $A(da)^2 + 2B(da)(db) + C(db)^2$:

$$A_{0,2} = \text{coeff}\,(b^2(da)^2) = -\cos^2 2\varphi, \quad C_{2,0} = \text{coeff}\,(a^2(db)^2) = -\cos^2 2\varphi,$$

$$2B = \text{coeff}\,((da)(db)) = 2ab(1+\cos^2 2\varphi) \text{ and } 2B_{1,1} = \text{coeff}\,(ab) = 1+\cos^2 2\varphi.$$

The formula of Lemma 1 provides the value of the curvature:

$$\begin{aligned} K &= 2B_{1,1} - A_{0,2} - C_{2,0} \\ &= 1 + \cos^2 2\varphi + \cos^2 2\varphi + \cos^2 2\varphi \\ &= 1 + 3\cos^2 2\varphi. \end{aligned}$$

Theorem 1 is proved (for the anomaly $\alpha = 2\varphi$). □

9.6 Ricci curvature and scalar curvature

The standard terminology in the studies of the Riemann tensor uses also the following numbers.

Definition. The *Ricci curvature* $R(\xi)$ in the direction of the tangent vector $\xi \in T_xM^n$ is the mean value of the sectional curvatures in all the 2-dimensional planes $V \subset T_xM^n$, containing this direction.

Remark. Fixing the unit vector ξ and choosing as η its complementary unit vector in the 2-plane V, we obtain (for the different choices of V) the sphere $\mathbb{S}^{n-2}$ of possible vectors η. The value

$$K(\xi, \eta) = \langle\, \Omega(\xi, \eta)\, \xi\,,\, \eta\,\rangle$$

is a quadratic form of the vector η.

Lemma. *The mean value of the quadratic form $\langle A\eta, \eta\rangle$ on the unit sphere $\langle \eta, \eta\rangle = 1$ of a Euclidean vector-space W^m is equal to* $(\mathrm{tr}A)/m$.

Proof. The mean value is a linear function with respect to A. It is invariant under the rotations of the Euclidean space W^n. The quadratic form is reducible, by a rotation, to the sum of the simplest forms $\sum a_k x_k^2$. Thus the mean value is the combination $\sum a_k m_k$ of the mean values m_k of the simplest forms x_k^2. To find these numbers m_k, note that they are all equal, since there are suitable rotations transforming one form to the other, and note that the quadratic form $\langle x, x\rangle = \sum x_k^2$ has the (mean) value 1 at the sphere $\langle x, x\rangle = 1$. Thus $m_k = 1/m$, and hence the mean value of the form $\langle A\eta, \eta\rangle$ along the sphere $\langle \eta, \eta\rangle = 1$ is $\sum(a_k/m)$, where the a_k's are the eigenvalues of the form $\langle A\eta, \eta\rangle$. Thus, $\sum a_k = \mathrm{tr}A$, proving the Lemma. □

Applying this lemma to the quadratic form $K(\xi, \eta)$ of the vector $\eta \in W^{n-1}$ (where the Euclidean vector-space W^{n-1} is the orthocomplement to the vector ξ in the tangent space T_xM^n), we obtain the formula

$$K(\xi, \eta) = \frac{\mathrm{tr}(\Omega|_{W^{n-1}})}{n-1},$$

where for ξ equal to the basic vector e_i, the numerator is the sum of the diagonal elements of the matrix of the linear operator $\Omega|_{W^{n-1}}$, that is,

$$R(e_i) = \frac{\sum_{j=1}^{n-1} R_{i,j,i,j}}{n-1}.$$

The anti-geometers, ignoring the preceding lemma, prefer to call by Ricci curvature rather the trace in the numerator of this fraction, instead of the mean value of the sectional curvatures. Thus, they obtain the strange value $(n-1)/r^2$ for the Ricci curvature of the sphere of radius r. We prefer the mean value specially for the applications to the infinite dimensional case, $n \to \infty$ (discussed below) which are important for hydrodynamics and for the mathematical theory of weather prediction.

The Ricci curvature $R(\xi)$ is a quadratic form of the tangent vector $\xi \in T_xM^n$.

Definition. The *scalar curvature* R of a Riemannian manifold M^n at a point x, is the mean value of the quadratic form $R(\xi)$ along the unit sphere $\langle \xi, \xi\rangle = 1$ in T_xM^n.

As in the case of the Ricci curvature, instead of the mean value of the mean value of a quadratic form, the anti-geometers prefer the trace of the matrix of the quadratic form (which is n times larger). So they call "scalar curvature" the number

$$R = \sum_{i=1}^{n} \sum_{j=1}^{n-1} R_{i,j,i,j},$$

which is $n(n-1)$ times bigger than the mean value.

We prefer to have the scalar curvature $R = 1/r^2$ for a sphere $\mathbb{S}^n$ of radius r.

Problem. Calculate the Ricci curvatures and the scalar curvature of 1. The spheres $\mathbb{S}^n$ of radius r; 2. The Lobachevsky spaces; 3. The complex projective spaces; 4. The unitary groups (equipped with the Riemannian metric invariant under the left and right diffeomorphisms L_g and R_g); and 5. The orthogonal groups (with their bi-invariant metric).

The sectional curvature of a Riemannian manifold provides a very important information on the behaviour of the geodesic lines on it.

The theorem that we will formulate below is more or less equivalent to the definition of sectional curvature (although it had been discovered and used by Jacobi earlier than the Riemannian curvature appeared).

Let γ be a geodesic line of a Riemannian manifold, parametrised by the length s of its segment from some initial point – Fig. 9.6.

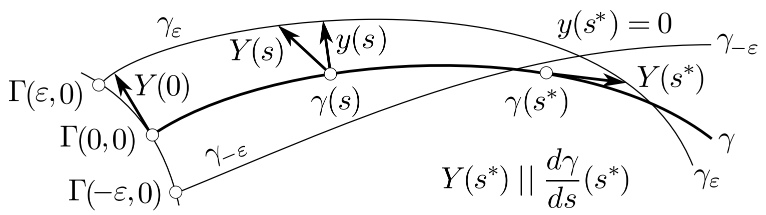


Figure 9.6: The normal variation field $y(s)$ along a geodesic line γ that is perturbed to become γ_ε.

Consider a neighbouring geodesic line. We associate to an (infinitesimal) deformation of γ, a normal vector-field called the *normal variation*,

$$y(s) \in T_{\gamma(s)}M^n \quad (\text{with } \langle y, d\gamma/ds \rangle = 0),$$

which is the normal component of the full variation vector

$$Y(s) = \frac{d}{ds}|_{\varepsilon=0}\Gamma(\varepsilon, s),$$

where $\Gamma(\varepsilon, s) \in M^n$ is a family of geodesic lines $\gamma_\varepsilon(s) = \Gamma(\varepsilon, s)$ of M^n. This family is parametrised smoothly by ε and reduced to γ for $\varepsilon = 0$.

Theorem 2. *The normal variation field verifies the following Jacobi* (second order differential) *equation:*

$$\frac{D^2 y}{(Ds)^2} = -\Omega\left(\frac{d\gamma}{ds}, y\right)\left(\frac{d\gamma}{ds}\right).$$

The "covariant derivative" D in the left hand side is defined in the following way: One transports back the vector $y_{(s_0+\delta)}$ from the point $\gamma(s_0+\delta)$ to the point $\gamma(s_0)$ (by a parallel transport along the geodesic γ), and then one differentiates the transported vector (considered as a function of the parameter δ with values in the fixed vector space $T_{\gamma(s_0)}M^n$) with respect to the "time" δ.

In the 2-dimensional case the normal variation may be understood as a scalar function, meaning the distance to the infinitesimally perturbed geodesic.

Example. For the flat case, $K = 0$, the Jacobi equation

$$\frac{D^2 y}{(Ds)^2} = 0$$

means the linear growth $y = c_0 + c_1 s$ of the distance between the neighbouring geodesic (straight) lines – Fig. 9.7.

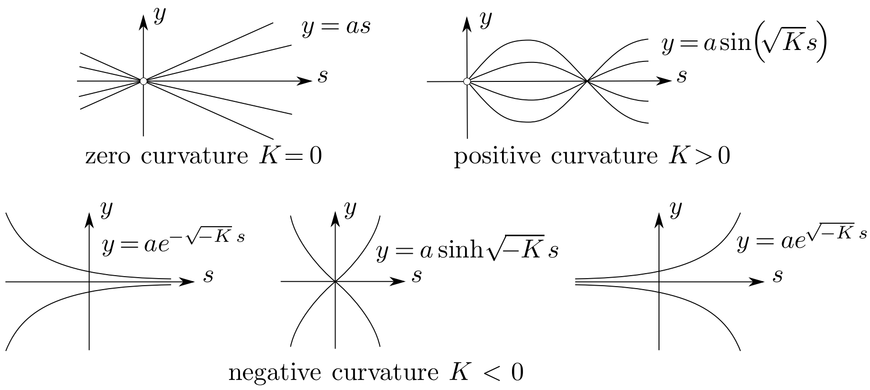


Figure 9.7: Behaviour of the normal variations $y(s)$ along a geodesic γ, for the sectional curvatures $K = 0$, $K > 0$ and $K < 0$.

For the case of the sphere, $K = r^2 > 0$, the Jacobi equation $\frac{D^2 y}{(Ds)^2} = -Ky$ describes the oscillatory behaviour of the normal variation (representing, say, the neighbouring meridian distance from the unperturbed meridian γ at the point $\gamma(s)$).

For the case of constant negative curvature, $K < 0$, the Jacobi equation $\frac{D^2y}{(Ds)^2} = -Ky$ describes the variation of the Lobachevsky straight lines. Say, in the notations of the Poincaré model, for the transition from the line $b(s) = 0$, $a(s) = e^s$ to line $b(s) = \varepsilon$, $a(s) = e^s$ the normal variation is $y(s) = e^s$, while the normal variation $y(s) = e^{-s}$ corresponds to a geodesic line intersecting the absolute curve at 0 (represented by an Euclidean circle, orthogonal to the absolute).

9.6.1 Conjugate Points and Caustics

The zeros of the solutions of the Jacobi equation, verifying the initial condition $y(0) = 0$, represent the intersections of the neighbouring geodesic lines (for instance, the meridians of the sphere starting at the north pole meet at the south pole).

These points of M^n are called the *conjugate points* of the original point x along the corresponding geodesic starting at x (Fig. 9.8). We see from the Jacobi equation that there are no conjugated points on surfaces with everywhere negative curvature.

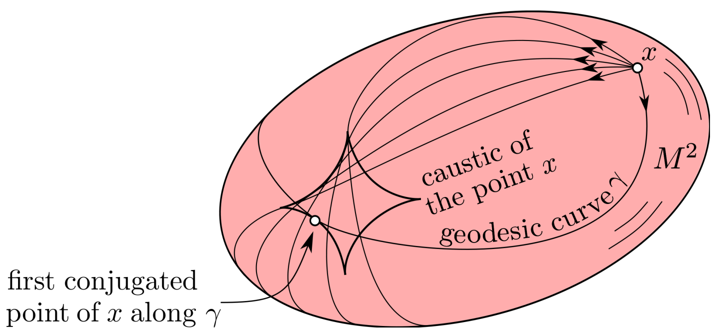


Figure 9.8: First conjugated points of the point x of a surface M^2 along different geodesic lines γ of M^2, starting at x.

The only conjugated points of a given point on the standard sphere are the point itself and the opposite point.

Perturbing the standard metric (say, replacing the sphere by a slightly different ellipsoid), Jacobi observed that the set of conjugated points of a given point (along all the geodesics starting at that point) is generically a closed curve (taking the first conjugated point, with minimal s, for each geodesic) – see Fig.9.8. This curve is called the *caustic* of the initial point x.

For the surface obtained from the sphere by a slight perturbation, the caustic of the (perturbed) point x (say, the North pole) is a small curve situated in a small neighbourhood of the South pole of the initial sphere.

The second conjugate points form a second closed caustic, and so on.

Jabobi proved that the caustic can't be smooth – there are, generically, cusp points. Jacobi proved that the number of cusps is even, announcing also that the caustic of a point of the ellipsoid has exactly four cusps.

It is known today that on any surface, not too far from the standard sphere, *the first caustic has at least four cusps.* Since is too difficult to prove here this topological theorem of contact geometry, we shall only mention that it is closely related to the following theorem on real Fourier series of functions on the circle (Sturm, 1836 [120]; Hurwitz, 1903 [85]) :

Theorem. *Let f be a 2π-periodic smooth function such that*

$$f(x) = \sum_{k \geqslant n} (a_k \cos(kx) + b_k \sin(kx)).$$

Then f has at least $2n$ zeros on its period $0 \leqslant x < 2\pi$.
(as it is the case for its first Fourier harmonics with non zero coefficient)

The classical 4-cusp theorem is related to the case $n = 2$ (see p. 617).

9.6.2 Negative Curvature and Unstability

The Riemannian manifolds with negative sectional curvatures have highly non stable geodesic lines. Indeed, the growth of the solutions of the Jacobi equation —like $e^{\sqrt{-K}s}$, where K is the curvature— implies that a small perturbation ε of the initial point or of the initial direction of the geodesic line, leads to a new geodesic whose distance from the original one grows exponentially, like $\varepsilon e^{\sqrt{-K}s}$, along the original geodesic.

Remarks. Literally, it is true only for sufficiently small $\varepsilon < \varepsilon_0(s)$, since when $\varepsilon e^{\sqrt{-K}s}$ is not small, the Jacobi linearised equation does not provide an adequate description of the distance (the two geodesics are no longer neighbouring).

The words "exponentially large distance" mean "exponentially large with respect to the original distance". The theory of chaotic dynamical systems proves that such an exponential growing of the distances provides the chaotic behaviour of the orbits of the dynamical system (of the geodesic lines in our case), making the "future" practically unpredictable when the "time" s becomes large with respect to $1/\sqrt{-K}$: It is the principle "small causes, big consequences", basic in statistical physics and in ergodic theory of chaotic behaviour.

Discussing the growth of the distance, we have written above the expression $e^{\sqrt{-K}s}$, which is the solution of the Jacobi equation only for the case of constant curvature K. In the general case, when K is not constant, there is no such simple solution of the Jacobi equation, but there are theorems evaluating the growth rate in terms of the value of the curvature $K(s)$. If

$$K(\xi, \eta) \leqslant -\kappa$$

for some constant $\kappa > 0$, then the generic solutions of the Jacobi equation are growing like the exponential $e^{\sqrt{\kappa}s}$ or even faster, providing the unpredictability for time intervals, large with respect to $\kappa^{-1/2}$.

9.6.3 Rigid body rotations and ideal fluids

The following (easy) theorems relate the geodesics of some interesting Riemannian manifolds to some problems in natural sciences.

Theorem 3 (Euler)**.** *The free rotations* $\{g(t)\}$ *of a rigid body* (say, a space research vessel or an asteroid) *form a geodesic line of the Lie group* SO(3)*, equipped with a left-invariant Riemannian metric.*

Proof. The kinetic energy of this motion along the rotation group is left-invariant, since the left multiplication of the motion $\{g(t)\}$ by a rotation h transforms it into the motion $\{hg(t)\}$, which is equivalent to the description of the original motion with respect to a different fixed coordinate frame.

Such a change of the coordinate system does not change the value of the kinetic energy, and hence the Lagrange function of our mechanical problem is left-invariant. According to the relativity principle, the geodesics are transformed to each other by the left translations. Theorem 3 is proved. □

Let M be a Riemannian manifold, and consider the group SDiff M of the volume-preserving diffeomorphisms of M, equipped with a right-invariant metric provided by the kinetic energy given by the quadratic form

$$H = \frac{1}{2}\int_M v^2(x)\tau,$$

where τ is the volume element of M, $v(x)$ is the velocity vector of the fluid particle at x and $v^2 = \langle v, v\rangle$ is the scalar square of v, determined by the Riemannian structure of M.

Theorem 4 (Euler)**.** *The motion* $\{g(t)\}$ *of an ideal* (incompressible, unviscous) *fluid filling the Riemannian manifold* M*, is a geodesic line of the group of volume-preserving diffeomorphisms* SDiff M*, equipped with the Riemannian metric given by the kinetic energy.*

Proof. The kinetic energy H is right-invariant because the right multiplication of the motion $\{g(t)\}$ by an element h of the group SDiff M transforms $\{g(t)\}$ into the motion $\{g(t)h\}$, which is equivalent to the description of the original motion, differing only by the renumbering h of the particles forming the fluid.

The incompressibility of the fluid and the preservation of the volume element τ by the renumbering $h \in \mathrm{SDiff}\, M$ implies the total independence of the kinetic energy on the renumbering (like $a_1+a_2+a_3 = a_2+a_1+a_3$). Thus we have proved the right invariance of the kinetic energy quadratic forms H on the vector spaces $T_g(\mathrm{SDiff}\, M)$ tangent to the infinite dimensional Lie group $G = \mathrm{SDiff}\, M$ at its points g. □

The hydrodynamic description of the motion of an ideal fluid as a geodesic of the kinetic energy metric, is one of the standard definitions of the ideal fluids and of the Euler equation of their motion.

It follows, from Theorems 3 and 4, that to investigate the behaviour of the motions of rigid bodies and of ideal fluids (under small perturbations of the initial conditions) one should study, respectively, the sectional curvatures of the left-invariant and right-invariant metrics on the manifolds of the corresponding Lie groups.

The explicit formulas for these curvatures are written below. They originated in [4], and a modern discussion can be read in the book [41].

To describe the left-invariant metric on a Lie group G, we start from the scalar product $\langle\ ,\ \rangle$, which defines a Euclidean structure on its Lie algebra $T_eG = \mathfrak{g}$. We define a bilinear map $B : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ by the formula

$$\langle B(a, b), c\rangle = \langle a, [b, c]\rangle \quad \text{for any} \quad c \in \mathfrak{g}.$$

Theorem 5. *The sectional curvature of a group G (equipped with a left-invariant metric defining a scalar product $\langle\ ,\ \rangle$ on $\mathfrak{g}$), for the section defined by two orthonormal vectors $\xi, \eta \in \mathfrak{g}$, is given by the formula*

$$K(\xi,\eta) = \langle\delta,\delta\rangle + 2\langle\alpha,\beta\rangle - 3\langle\alpha,\alpha\rangle - 4\langle B_\xi, B_\eta\rangle,$$

where

$$2\alpha = [\xi,\eta], \qquad 2\beta = B(\xi,\eta) - B(\eta,\xi), \qquad 2\delta = B(\xi,\eta) + B(\eta,\xi),$$

$$2B_\xi = B(\xi,\xi), \quad 2B_\eta = B(\eta,\eta).$$

For the right-invariant metric the formula has the same form because the transition from "left" to "right" is equivalent to the change of signs of the operations $[\ ,\ ]$ and B, which leaves K unchanged.

Applying this theorem to the group $\mathrm{S\,Diff}\ \mathbb{T}^2$ of area-preserving diffeomorphisms of the torus with a plane metric, we shall represent two divergence-free vector fields ξ and η of $\mathfrak{g}$ by their respective Hamiltonian functions (that we note α and β) called in hydrodynamics "flux functions".

Example. Consider two "wave vectors" a and b of a Euclidean plane covering the torus, and consider the flux functions

$$\alpha(p) = \cos\langle a,p\rangle, \quad \beta(p) = \cos\langle b,p\rangle,$$

where p is a point of the torus $\mathbb{R}^2/(\sigma\mathbb{Z}+\tau\mathbb{Z})$, for some suitable period vectors σ and τ such that the numbers $\langle a,\sigma\rangle$, $\langle a,\tau\rangle$, $\langle b,\sigma\rangle$ and $\langle b,\tau\rangle$ are integer multiples of 2π.

For the Hamiltonian vector fields ξ and η corresponding to the flux functions α and β, the preceding theorem provides the sectional curvature

$$K(\xi,\eta) = -\frac{a^2+b^2}{4S}\sin^2(\varphi)\sin^2(\psi),$$

where S is the area of the torus, φ is the angle between the vectors a and b, and ψ –between the vectors $a+b$ and $a-b$ of the covering Euclidean plane.

For the torus $\mathbb{T}^2 = \{x \pmod{2\pi}, y \pmod{2\pi}\}$ the sectional curvature of the group $\mathrm{SDiff}\mathbb{T}^2$, in the direction defined by the flows of the vector fields $\sin y(\partial/\partial x)$ and $\sin x(\partial/\partial y)$, equals $K = -1/(8\pi^2)$.

There exist sections with even more negative curvatures.

Unstability and weather prediction. The negativeness of such sectional curvatures is the principal reason for the impossibility of the dynamical prediction of the weather. Namely, a perturbation of the initial velocity field will grow approximately 10^5 times in the period of one month, making impossible to predict the motion of the atmospheric mass for such a long time. Indeed, an initial displacement of 1 Km would grow in a month approximately to the scale of the whole planet ($\approx 10^5$ Km).

Even the irregular displacements or velocity disturbances, whose averages along each square kilometre of the Earth surface are zero and which no receptors would ever be able to detect at the initial moment, will produce a change of the weather at the planetary scale, in few weeks.

Of course, the meteorologists prefer to promise more, in order to get the founding for the monstrous computations, expecting the miraculous capability to compute the weather a season ahead. We hope this progress of computational meteorology will provide useful results, both for the computer scientists and for its users. But the calculation of the curvatures (of which we have quoted above the first simplest examples) indicates that no computational progress would ever make the predictions of dynamical weather reliable for any period longer than few weeks (10 days?).

9.7 Poincaré series of classification problems

Besides the theory of the classification of Riemannian metrics, started by Riemann, we shall mention a more general theory —elaborated by his followers, especially by E. Cartan— which as far as we know, remains unpublished in its most general form.

In this theory one considers *any local classification problem.* However, the explicit description of the above word "any" is still missing in this general theory.

Examples. The classifications up to local diffeomorphism of the smooth (or analytic or holomorphic) manifolds, or of the smooth functions defined in a neighbourhood of 0, or of the smooth vector fields, or of the smooth Riemannian metrics, or of those "Einstein metrics" whose Ricci curvature vanishes identically, or of the solutions of a system of partial differential equations (in a neighbourhood of 0).

Sometimes the classification is finite, like for the functions at a non degenerate critical point. In other cases it is not, and there may exist continuous invariants of the objects under study, whose values cannot be changed by a choice of the coordinate system, like the eigenvalues of the linearisation of a vector field at a point where it vanishes.

Sometimes one needs "functional moduli" (that is, a set of arbitrary functions) to parametrise the different objects of our classification. A typical example is provided by the boundary conditions, or by the initial conditions, which determine a particular solution among all the solutions of a given differential equation.

The problem is to provide an exact definition of these "necessary arbitrary functions parametrising the non equivalent objects of the classification". We shall describe an approach to this general problem, by developing the Riemann approach to the local classification of the Riemannian metrics, up to diffeomorphisms (which led him to the discovery of his tensor $R_{i,j,k,\ell}$).

The first idea is to approximate the infinite-dimensional space of our objects by its finite-dimensional approximations J^k (which consists of the Taylor polynomials of degree k of our objects).

The diffeomorphisms act naturally on this space, as a finite dimensional Lie group. So, the space J^k is decomposed into the orbits of this action, which are the equivalence classes of the k-Taylor series.

Write n_k for the dimension of the space of these orbits – it is the "moduli number" for the Taylor polynomials of degree k. To calculate the dependence of n_k on the degree k of the Taylor polynomial, we prefer to study the difference $m_k = n_k - n_{k-1}$, interpreting it as the "number of *new moduli* obtained from the terms of degree k" (supposing that the n_{k-1} parameters distinguishing the Taylor polynomial of degree $k-1$, would remain necessary for the classification of the more precise Taylor polynomials of degree k).

Definition. The *Poincaré series* of the sequence m_k is defined by

$$p(t) = \sum_{k=0}^{\infty} m_k t^k.$$

Remark. The series p should not be considered as a function. Here t is a formal variable and $p(t)$ is a formal series. What is important and defines the "series" is the sequence of coefficients $\{m_k\}$, which are just labeled with the respective powers $\{t^k\}$.

Example. Suppose our classification of objects depends on two (arbitrary) functions of one variable. In our notation it means that $m_k = 2$ for every k. Hence, the Poincaré series is $p(t) = \frac{2}{1-t}$.

Example. Suppose our classification invariant is an arbitrary even function of one variable. It means that $m_{2k} = 1$ and $m_{2k-1} = 0$. Hence, the Poincaré series is $p(t) = \frac{1}{1-t^2}$.

Conjecture. *The Poincaré series of the "natural problems" are rational functions.*

The difficult point here is the exact definition of "natural problem". We suggest to include at least the problems listed above: Local classification of vector fields, of exterior differential forms, of tensor fields, of quadratic differential forms, of systems of differential equations, of solutions of systems of differential equations, ...

One should also be careful with the definition of the numbers m_k or n_k because the orbit space is not always a manifold and has different dimensions at different places.

One may take the maximal dimension for all the Taylor series of a given degree, but there is a more attractive way. Namely, one may choose a particular object f, and consider the local dimension of the space of orbits at that point, that is, the number of moduli needed to parametrise the Taylor series of the neighbouring objects of f.

In this way one obtains a Poincaré series which depends on f, but the above conjecture has meaning in this case too. It may happen, for instance, that the rationality of the Poincaré series of f take place almost always (so that the exceptional objects f belong to a set of infinite codimension in the space of the infinite Taylor series of our objects).

The rationality conjecture is related to the idea of the parametrisation of arbitrary functions, since Euler proved the following basic theorem.

Theorem 6. *The Poincaré series of the space of (arbitrary) smooth functions of r variables (with no equivalences) is*

$$p_r(t) = \frac{1}{(1-t)^r}.$$

Example. The number of different monomials of degree k in 2 variables is $m_k = k+1$. Thus, $m_k - m_{k-1} = 1$, which means that $(1-t)p_2(t) = \sum t^k$. Knowing the last series to be $p_1(t) = 1/(1-t)$, we find $p_2(t) = \frac{1}{(1-t)^2}$.

Euler's theorem can be proved inductively by the same reasoning,

$$(1-t)p_r(t) = p_{r-1}(t).$$

Euler himself used more direct arguments. He multiplied the r series

$$\frac{1}{1-x} = 1 + x + x^2 + \ldots, \quad \frac{1}{1-y} = 1 + y + y^2 + \ldots,$$

for r different arguments.

The product

$$(1+x+x^2+\ldots)(1+y+y^2+\ldots)\ldots = 1+(x+y+\ldots)+(x^2+xy+y^2+\ldots)+\ldots$$

contains each monomial (of any degree) of our r arguments exactly once. Replacing all the arguments by t, we get the statement of Euler's theorem, since the number of repetitions of the monomial t^k at the right hand side will be equal to the number of monomials of degree k in r variables.

So, whenever the parametrisation of the equivalence classes is provided by some arbitrary functions (including even such cases as the case of the even functions, which are arbitrary functions of the argument x^2), the Poincaré series of the moduli space is given by some sum of series similar to $\frac{t^q}{(1-t^a)^b}$, which is a rational function.

A. Tresse, a student of Sophus Lie, formulated a general theorem describing the invariants of the "generic" analytic classification problems, in such a way that no modern mathematician has later explained or disproved. His theory is some kind of generalisation of the Hilbert theorem on the finite bases for the ideals in the rings of polynomials or of power series.

The invariants of a generic classification problem of analysis form a ring: The sum and the product of two invariants are also invariants.

Tresse considered also the "semi-invariants", whose vanishing varieties are invariantly related to the classification *(consist of full orbits)*. A semi-invariant represents an invariant variety by the ideal of the functions vanishing on this variety (in the ring of functions of the space of Taylor series).

Tresse called *invariant vector fields* those vector fields that send orbits (or ideals) to orbits (ideals).

These fields form a module over the ring of invariants (remaining invariant even when they are multiplied by any invariant function), and the derivative of an invariant (semi-invariant) along an invariant field is still an invariant (semi-invariant). The Poisson bracket of invariant fields is still an invariant field.

Tresse announced that "for any natural problem" all this algebra is generated by a finite number of invariants and of invariant fields: The invariant field multiplied from the left by an invariant function is still an invariant field. The multiplication by a function staying at right, means the derivation of the function along the field, and sends an invariant function to another invariant function. The product of two invariant fields is understood as their Poisson bracket and is a new invariant field.

Multiplying them by each other one obtain new invariants and new invariant fields whose sums provide, according to Tresse, all of them.

He calculated everything explicitly in the local classification problem of the second order ordinary differential equations,

$$\frac{d^2y}{dx^2} = F\left(x, y, \frac{dy}{dx}\right).$$

This problem of differential equations is equivalent to the local classification of the pairs of direction fields in the 3-space.

To understand the equivalence, we interpret one of the direction fields as the field of tangent directions to the fibres of the linear projection of the 3-space onto the 2-plane.

The projections of the integral curves of the second direction field to this plane define a 2-parameter family of smooth curves. The graphs of the solutions of the second order differential equations also form a 2-parameter family of curves (in the plane of the independent and dependent variables x and y).

In this way we (locally) interpret the pair of direction fields as a second order differential equation. The classification problem, considered by Tresse, was the local study of different 2-parameter families of smooth curves in the plane. For instance, which families are reducible to the family of the straight lines? That is, which second order differential equations can be locally reduced to the linear form $d^2y/dx^2 = 0$ by local diffeomorphisms of the plane with coordinates x and y?

The strange answer is that the right hand side F should be a cubic polynomial with respect to the 3rd variable, $p = dy/dx$. Tresse studied the whole ring of invariants in this problem, calculating the finite set of few invariants and invariant fields which generate all the invariants and all the invariant fields, whose Poincaré series were also calculated by him (proving their rationality).

A detailed description of this classification problem is given in the book [15], § 6, pp. 43-58. The forgotten papers of Tresse are [124] and [125].

For the problem of the local classification of the Riemannian metrics, the Poincaré series was computed by Shmelev in [116] (containing also the Poincaré series for the classifications of the Kählerian and of the hyperkählerian metrics, which are the complex and quaternionic sisters of the Riemannian metrics). The Poincaré series obtained by Shmelev are rational functions. His paper contained also the proof of the finiteness of the number of generating differential invariants, in the sense of Tresse, for these three classification problems.

Essentially, these results should provide a complete set of identities involving the Riemann tensor: One describes the basic identities from which all the others are differential algebraic linear combinations, one describes all *syzygies* (relations between these identities taking into account the dependences of the combinations), *second syzygies* (measuring the dependences of the preceding dependences), and so on. As far as we understand, all these spaces should have rational Poincaré series. It would be interesting to publish those series explicitly, at least for the three Shmelev classification problems.

For the Riemannian geometry we shall mention here only some few of the first identities (of the natural generators of the corresponding modules of the differential relations):

$$\Omega(\xi,\eta)\zeta+\Omega(\eta,\zeta)\xi+\Omega(\zeta,\xi)\eta=0, \tag{1}$$

$$\langle\Omega(\xi,\eta)\alpha,\beta\rangle=\langle\Omega(\alpha,\beta)\xi,\eta\rangle, \tag{2}$$

$$\Omega(\xi,\eta)\zeta=-\frac{D}{D\widetilde{\xi}}\frac{D}{D\widetilde{\eta}}\widetilde{\zeta}+\frac{D}{D\widetilde{\eta}}\frac{D}{D\widetilde{\xi}}\widetilde{\zeta}+\frac{D}{D\{\widetilde{\xi},\widetilde{\eta}\}}\widetilde{\zeta}, \tag{3}$$

where $\widetilde{\xi}$, $\widetilde{\eta}$ and $\widetilde{\zeta}$ are any continuations of the vectors ξ, η and ζ of T_xM, respectively, to smooth vector fields. The symbol $\frac{Dw}{Dv}$ (where v and w are two vector fields) means the covariant derivative of the vector field w along the flow of the field v. The meaning of the covariant derivative D was explained above, on page 343, where we used it to formulate the Jacobi theorem on normal variations.

It is a remarkable fact that the right hand side of formula (3) is independent of the continuations of the three vectors ξ, η, ζ. Note also that applying any differential polynomial to an identity we obtain a new identity: The Poincaré series discussed above indicate which of these identities are new, and which are corollaries of the preceding ones.

The attempts, at present day, to study and to continue the "Theory of non holonomic systems" created by E. Cartan (who hoped to generalise the works of Tresse), follow the ideology of algebraic geometry. So, people try to prove "always-true theorems".

An attracting example is Hilbert's finiteness basis theory, which is simultaneously true both in the generic (non degenerate) situation, as in the the implicit function theorem, and also for arbitrary polynomial ideals, independently of their level of degeneration.

An other attractive example is the theorem on the "resolution of singularities", claiming that every singular algebraic variety is the image of some non singular one, projected algebraically on the singular variety. The semi-cubic parabola $y^2 = x^3$ is the image of $\mathbb{C}$ for the map $y = t^3$, $x = t^2$, and the singularity resolution extends this construction to arbitrary algebraic varieties with any singularities.

Similarly, in the theories of partial differential equations of non holonomic systems, the experts are trying to extend the (natural) studies of the non degenerate situations to *all the algebraic degenerations simultaneously*, leaving the study of the degenerations hierarchy to the disdained calculators.

In this way many interesting things remain unobserved and unproved (we quoted above a small part of such problems and conjectures, like the definition of the "natural analytic classification problems", the Poincaré series conjecture and their explicit calculations and interpretations in practical important cases).

This algebraist ideology, neglecting the genericity analysis and restricting the theories to those results which do not need any genericity restrictions, would produce a lot of harm for mathematics if it were followed by everybody (which, happily, had not been the case in the past centuries).

Thus the rational numbers would be algebraically rejected, $\sqrt{2}$ being irrational. The field of real numbers would be rejected, being not closed algebraically. The theories of degree of smooth maps, of indices of vector fields and so on would be rejected as well as some exceptional polynomials of degree n having less that n (complex) roots.

Archimedes's and Newton's integration would be rejected by algebraic geometry, as the function $1/x$ which has no algebraic primitive function. Differential equations like $dx/dt = x$ would be unsolvable. Most parts of interesting mathematics would be rejected if the mathematicians were obliged to restrict themselves to those particular results which remain true for all the degenerated cases.

The hierarchy of the degeneracies in the "algebraic" theory of the non holonomic systems is still missing. Neither the degeneracies of smallest codimensions (which are generic in families depending in a small number of parameters), nor the "simple degeneracies" (where there are not continuous moduli) have been classified in this theory, in spite of many papers on the "algorithms to find the general solutions" of non holonomic systems.

A celebrated success of the algebraist ideology was the solution of this problem : What is the maximal length of a bar which one can transport along the tunnel of Fig. 9.9 ?

The celebrated Tarski-Zeidenberg theorem of general algebraic geometry claims that the algebraic projection of any algebraic or semialgebraic set (defined by algebraic equa-

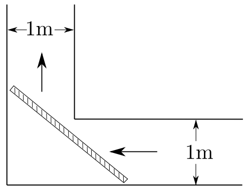


Figure 9.9: The problem of passing a bar along a "right angled" tunnel.

tions and inequalities) is itself a semialgebraic set.

It follows that the problem of the maximal bar is semialgebraic, and hence that there exists a finite algebraic algorithm for its solution. The computerised realisation of this general algorithm for the particular problem of Fig. 9.9 occupied several hundred of pages. So, the computer calculated the answer in few days.

For every natural scientist it is geometrically evident that the maximal length equals $2\sqrt{2}$. But the general ideology from algebraic geometry suggest a different way for mathematics, considered by them as the art of finite formal games with symbols.

Returning to the problem on the boundary and initial conditions of the partial differential equations, or to the more general Tresse program of the invariants of the natural classification problems, one might replace the algebraist approach by the hierarchical approach of Poincaré to bifurcations and singularity theory (see pp. 562-567 where this approach is discussed).

This program requires to distinguish first the generic objects, studying the complement to the "variety of degenerations" in the space of objects under study (differential equations, Riemannian metrics, boundary conditions, and so on).

For the non-degenerate cases one should provide finite algorithms and the classifications, similar to the Implicit Function Theorem or to the Cauchy-Kovalevskaya Theorem on PDE. For the degenerate cases one should provide the evaluation of their codimension and the finite description of the degenerations of small codimension. The next step should provide the study of these first degenerations (similar to the study of the generic case, which was the starting point).

And next, one continue in the same way: The secondary degenerations, violating the algebraic conditions needed for the study of the problem at the previous state, provide a new class of problems in which one have to find the most generic cases and the exceptional ones of the next level, and so on.

Unfortunately, these hierarchies of the degenerations (which one may also call *the natural stratification of the space of objects under study*) are still unknown even in the most natural classification problems, like the theory of non-holonomic systems, the boundary problems of PDE, the Riemannian metrics or the Einstein metrics.

Chapter 10

Degree, index and linking

We already know the fundamental group of the circle, $\pi_1(\mathbb{S}^1) \cong \mathbb{Z}$. We shall compute here the higher dimensional version of it, the group $\pi_n(\mathbb{S}^n) \cong \mathbb{Z}$, calculated by Poincaré, who used the previous ideas of Sturm and Kronecker. Sturm had invented the "Sturm characteristic" of two polynomials of one real variable, to evaluate the number of real roots of a polynomial on a segment by counting the "sign changes" in some sequences. Kronecker had extended it to three real polynomials of two variables, or to $n+1$ polynomials of n variables.

Kronecker claimed that God had created only the natural numbers, and that all other mathematical objects are human inventions. So, for him arithmetics and number theory were the most important parts of mathematics, and his studies of real polynomials were intended to the case of integer coefficients. But doing this arithmetics cleverly, he invented some extremely important mathematical notions that in modern mathematics are known as degree of maps, index of vector fields, intersection index and so on, which, in Poincaré hands, became homology theory, intersection (and cohomology) rings, linking theory and cobordism theory, as we shall see soon.

10.1 Degree of a map

We recall first the isomorphism between the fundamental group of the circle and the additive group $\mathbb{Z}$, provided by the *degree* of the corresponding loop, which is the number of its turns around the circle.

Of course, to obtain this degree isomorphism, we have to fix the orientations of the 1-dimensional cube $I = \{0 \leqslant t \leqslant 1\}$ and of the circle $\mathbb{S}^1$ to which the cube I is sent by the spheroid $\varphi : I \to \mathbb{S}^1$.

To define this "number of turns", we have used the angular coordinate ϑ on the circle, orienting it positively. Defining the angular coordinate $\vartheta(\varphi(t))$ continuously, we count its increment $\vartheta(\varphi(1)) - \vartheta(\varphi(0)) = 2\pi k$, and the integer k is the degree of the map φ (equal to the degree of any other spheroid map ψ homotopic to φ).

In the case of higher dimensions we miss the angular coordinate ϑ because the sphere $\mathbb{S}^n$ is simply connected: $\pi_1(\mathbb{S}^n) = 0$ for $n > 1$.

So, to extend the definition of degree to higher dimensions, we ought to replace its description in terms of the function $\vartheta(\varphi(\cdot))$ by a different counting of the number of turns.

The following definition provides clearly the same value of the degree for the case of the circle $\mathbb{S}^1$, but using no angular coordinate at all.

Since this definition will be also used for the case of higher dimensions, we shall formulate it in this way already now. To get the case of the fundamental group of the circle, one should simply take $n = 1$ in the following definition.

First, consider a smooth spheroid map,

$$\varphi : (I^n, \partial I^n) \to (\mathbb{S}^n, *),$$

where the cube $I^n = \{t \in \mathbb{R}^n : (0 \le t_1 \le 1, \dots, 0 \le t_n \le 1\}$ is oriented by the ordering of the oriented coordinates t_j $(1 \le j \le n)$, and the orientation of the sphere $\mathbb{S}^n$ is also fixed.

For any class of continuous spheroid there is a homotopic smooth representative φ of the same homotopy class $[\varphi]$. We shall work with this smooth representative, obtaining that the resulting degree would be independent of the choice of the smooth representative.

To count the "number of turns", we take some point ω of the sphere-image $\mathbb{S}^n$ and we count the number of points sent by φ to ω. Of course, this number may depend on ω, and ω could have an infinite number of preimage points.

To avoid the last difficulty, we choose for ω a non critical value of the smooth map φ, different from the value $*$. Such a value exists, since the set of critical values has measure zero in $\mathbb{S}^n$, by the Sard Lemma (pp.20-23).

For a non critical value ω the map φ is, by the implicit function theorem, a local diffeomorphism in the neighbourhood of any preimage point $x \in \varphi^{-1}(\omega)$, and hence the complete preimage set, $\varphi^{-1}(\omega) \subset I^n$, consists of a finite number of interior points of the cube (Fig. 10.1):

$$\varphi^{-1}(\omega) = \{x_1, \dots, x_m\}, \quad x_j \in I^n \setminus \partial I^n.$$

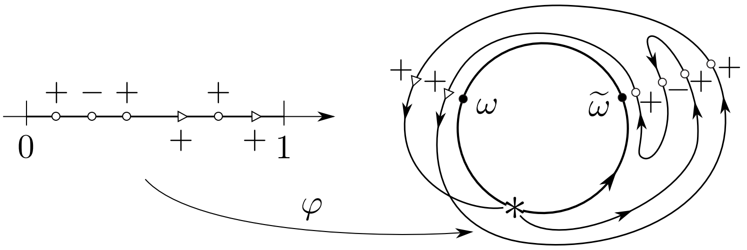


Figure 10.1: Signs of the preimages of the non critical values ω and $\widetilde{\omega}$ of a spheroid φ of degree 2.

To make the result of the counting independent of the choice of ω, we shall not count the "number of turns" by m (see Fig. 10.1). We associate a *sign* to each preimage point x_j, and then we define the "number of turns" as the sum of the m terms ± 1 associated to the m points x_j, where the sign of the points depends on the chosen orientations as follows.

The local diffeomorphism φ sends the orienting frame $(e_1, \ldots, e_n)$ of the cube I^n at x_j ($e_p \in T_{x_j} I^n$) to the point $\varphi(x_j) = \omega$ of the sphere, providing there some orienting frame $(f_1, \ldots, f_n) = (A_j e_1, \ldots, A_j e_n)$, where $A_j : T_{x_j} I^n \to T_{\varphi(x_j)} \mathbb{S}^n$ is the derivative of the local diffeomorphism φ at x_j.

The point x_j is *positive* if the frame $(f_1, \ldots, f_n)$ orients the sphere $\mathbb{S}^n$ positively, being *negative* in the opposite case (see Fig. 10.1).

Definition. *The degree of the spheroid* φ is the sum of the signs $\varepsilon_j \in \{+1, -1\}$ of the m points x_j of the preimage $\varphi^{-1}(\omega)$ of a non critical value ω:

$$\deg \varphi = \sum_{j=1}^{m} \varepsilon_j \in \mathbb{Z}.$$

We have to prove the independence of this integer from the choice of the non critical value ω.

In Fig. 10.1, we obtain $m = 2$ for the non-critical value ω and

$$\deg \varphi = 1 + 1 = 2.$$

For the other choice, $\widetilde{\omega}$, we get $\widetilde{m} = 4$, but again

$$\deg \varphi = 1 - 1 + 1 + 1 = 2.$$

To prove that this ω-independence holds in the general case too, we shall prove a more general statement: The number

$$\sum_{x\in\varphi^{-1}(\omega)} \varepsilon(x)$$

does not depend even on the particularities of the map φ, it is the same for all smooth maps homotopic to φ.

Among such homotopies, we have the families of orientation-preserving diffeomorphisms of the sphere which can move ω to $\widetilde{\omega}$. The independence of the sum of the signs (of the preimages) from the choice of the point ω, for such variable spheroid φ, provides the independence of the sum of the signs for the original spheroid φ, equipped with two different non critical values ω and $\widetilde{\omega}$.

So, suppose $\Phi : I^n \times I \to \mathbb{S}^n$ to be a smooth representative of the homotopy $\{\varphi_s\}$ between the spheroids

$$\varphi_0 : I^n \to \mathbb{S}^n \quad \text{and} \quad \varphi_1 : I^n \to \mathbb{S}^n,$$

where $\varphi_0(t) = \Phi(t,(s=0))$ and $\varphi_1(t) = \Phi(t,(s=1))$ for any $t \in I^n$.

To simplify the subsequent formulas, we suppose that $\varphi_s \equiv \varphi_0$ for some neighbourhood of $s = 0$ and that $\varphi_s \equiv \varphi_1$ for some neighbourhood of $s = 1$ in I.

Consider the $(n+1)$-cube $I^n \times I = \{t, s\}$ equipped with its standard Euclidean metric, whose square is $dt^2 + ds^2$.

Take a non critical value $\omega \in \mathbb{S}^n$ of the smooth map Φ, different from $*$. It is a non critical value also for the restrictions φ_0 and φ_1 of Φ.

By the implicit function theorem, the preimage $\gamma = \Phi^{-1}(\omega)$ is a smooth curve in the cube I^{n+1} – Fig. 10.2.

This curve γ may have several connected branches. Each branch is either a closed curve (an embedded circle) or a smooth segment ending at the faces $s = 0$ and $s = 1$, being orthogonal to these faces at the end-points.

Consider the Euclidean subspace V of dimension n, normal to the curve γ at its point p. The hyperplane V depends continuously on the point p of γ where we have constructed it. The derivative of the map Φ at p sends this n-space V to the tangent space $T_\omega\mathbb{S}^n$ isomorphically, and the preimage of an orienting frame $(f_1, \ldots, f_n)$ of this tangent space provides a frame $(e_1, \ldots, e_n)$ of the n-space V.

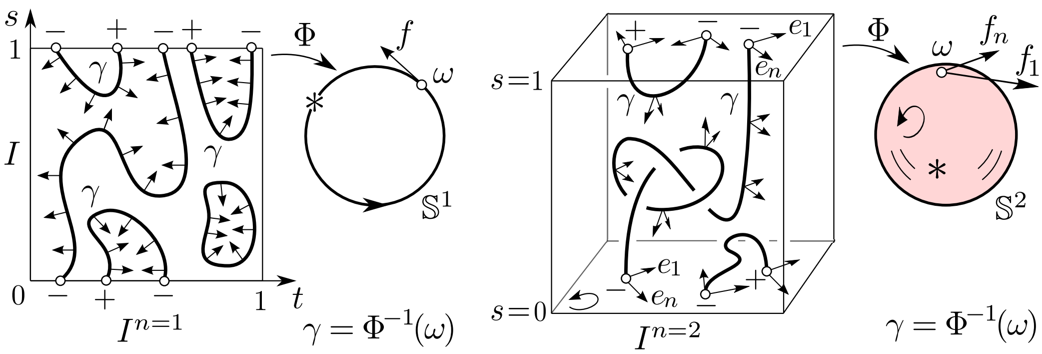


Figure 10.2: The framed preimage curve γ for the map $\Phi : I^2 \to \mathbb{S}^1$ (and $\Phi : I^3 \to \mathbb{S}^2$) defining the homotopy of the spheroids φ_0 and φ_1 of dimension $n = 1$ (and $n = 2$ respectively).

We choose a vector e_0 tangent to the curve γ and depending continuously on p in order to get a frame $(e_0, e_1, \ldots, e_n)$ that orients the $(n+1)$-cube I^{n+1} positively (for every point p) – Fig. 10.3.

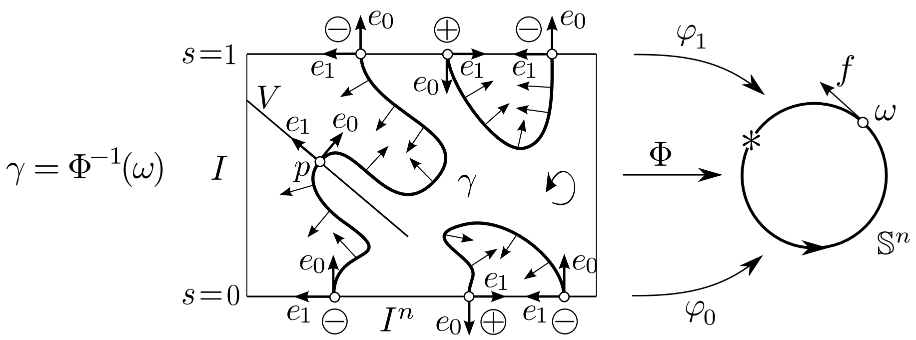


Figure 10.3: Proof of the degree invariance under the homotopy Φ from φ_0 to φ_1.

Main Lemma. *The vectors* $(e_1, \ldots, e_n)$ *orient the* n*-cube* I^n *at the two end-points of a component of the curve* γ *in opposite way if both end-points belong to the same face* ($s = 0$ or $s = 1$), *and in the same way if they are in different faces.*

Indeed, since the $(n + 1)$-frame orientation is fixed, the orientations of the vector e_0, normal to the face, are opposite at the two end-points in the case of return of γ to the initial face, and are positively parallel to the initial direction if the curve ends at the opposite face.

Corollary. *The sums of the signs of the preimages of the non critical value ω for both homotopic spheroids φ_0 and φ_1 are equal.*

Proof. Every component of the curve γ returning to the initial face, contributes nothing to the sums of the signs of the preimages of ω, both for the spheroid φ_0 and for the spheroid φ_1.

Every component connecting both sides contributes with summands of equal sign to both sums.

Hence,

$$\sum_{\varphi_0(x)=w} \operatorname{sign}(x, \varphi_0) = \sum_{\varphi_1(y)=w} \operatorname{sign}(y, \varphi_1).$$

□

Thus the degree is well defined as a map $\deg : \pi_n(\mathbb{S}^n) \to \mathbb{Z}$ and does not depend neither on the choice of the non critical value ω that we used to count it, nor on the choice of the representative φ of the homotopy class of the spheroid.

Claim. *The map* $\deg$ *is a group homomorphism.*

Proof. The preimage of a non critical value ω for the product spheroid $\varphi\varphi'$ is formed by the $m+m'$ preimages of ω in the union of two cubes (one sent to $\mathbb{S}^n$ by the spheroid φ and the other sent by φ'). Since the signs are defined in the same way, we have

$$\sum_{(\varphi\varphi')^{-1}\omega} \text{signs} = \sum_{\varphi^{-1}(\omega)} \text{signs} + \sum_{(\varphi')^{-1}(\omega)} \text{signs}\,,$$

proving the claim. □

The degree is the *only* homotopy invariant of n-spheroids on $\mathbb{S}^n$:

Theorem 1. *For any integer k there exists an n-spheroid of degree k on the sphere $\mathbb{S}^n$, and it is unique up to homotopy*:

$$\pi_n(\mathbb{S}^n, *) \cong \mathbb{Z}\,. \tag{1}$$

Corollary. *The real projective spaces satisfy* $\pi_n(\mathbb{R}\mathrm{P}^n) = \mathbb{Z}$. (See p. 77.)

Proof of Theorem 1. Decompose the image sphere into two “hemispheres”: a neighbourhood U of the non critical value ω and the remaining part $\widetilde{U} = \mathbb{S}^n \smallsetminus U$ (which is a closed disc).

This pair is diffeomorphic to the pair: (open North hemisphere, closed South hemisphere). So, there exists a homotopic contraction of the part $\widetilde{U}$ to the point $*$: each point of the sphere follows its “meridian”, and the family of homotopic maps starts from the identity map (at $s = 0$) and ends at the contracting map sending all the disc $\widetilde{U}$ to the point $*$, the whole complementary part to the point $*$ in the sphere, $\mathbb{S}^n \smallsetminus *$, is then covered once by each of the small n-discs U_j.

This homotopy, following the spheroid φ, transforms this spheroid homotopically to a new spheroid $\widetilde{\varphi}$ which sends diffeomorphically the initial small neighbourhoods $\varphi^{-1}(U)$ of the preimage points $x_j \in \varphi^{-1}(\omega)$ to $\mathbb{S}^n \smallsetminus *$, and the whole complement of the union of these m disjoint open n-discs in I^n is sent by the new spheroid to the prescribed point $* \in \mathbb{S}^n$ – Fig. 10.4.

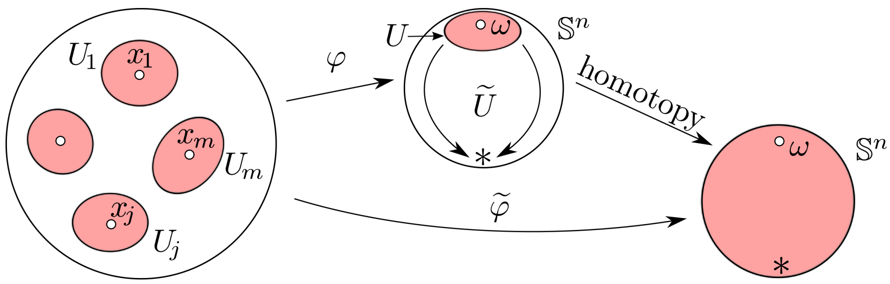


Figure 10.4: The special spheroid $\widetilde{\varphi}$ provided by an arbitrary spheroid φ.

We have thus proved an important proposition

Proposition. *Each n-spheroid of the n-sphere $\mathbb{S}^n$ is homotopic to a special n-spheroid which sends diffeomorphically to $\mathbb{S}^n \smallsetminus *$ each of some m disjoint small open n-discs of I^n, and sending to the point $*$ their whole complement.*

It is quite easy to construct such a map of any given degree (using for the negative degrees the reflected diffeomorphisms). This construction proves the uniqueness of the degree invariant and the isomorphism (1).

Indeed, any two maps of equal degree are homotopic to two of the special maps described above, which may be also chosen to have the same number of discs $m = |\deg \varphi|$ (using the evident homotopy $\varphi\varphi^{-1} \sim 0$ for the identity map φ).

Now the two spheroids of equal degree are homotopic to the same special spheroid of this degree, and hence are homotopic to one another, proving the degree uniqueness (1) of Theorem 1. □

Remark. Our reasoning proves more than formula (1). Indeed, the degree may be defined for the maps of any closed oriented n-manifold to the n-sphere,

$$\varphi : M^n \to \mathbb{S}^n,$$

by the same counting of the signed preimages in $\varphi^{-1}(\omega)$. The resulting integer, $\deg \varphi \in \mathbb{Z}$, is independent of the choice of the non critical value ω as well as of the choice of the representative φ of its homotopy class.

The proofs remain the same as for the spheroids case, except for the group operation of the spheroids, which is not defined for the maps of general manifolds M^n.

10.1.1 Degree of Polynomials and of Rational Functions

Consider a complex polynomial

$$f(z) = z^n + a_1 z^{n-1} + \ldots + a_n$$

as a smooth map $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$. Adding a point $z = \infty$ at infinity, we close the affine line $\mathbb{C} \approx \mathbb{R}^2$ to become the "Riemann sphere" $\mathbb{C}\mathrm{P}^1 \approx \mathbb{S}^2$.

Problem. Show that the map φ *becomes smooth in the neighbourhood of the "point at infinity"* $\infty = \varphi(\infty) \in \mathbb{C}\mathrm{P}^1$.

Solution. Using the coordinate $w = 1/z$ at the neighbourhood of $z = \infty$, we express φ in this neighbourhood in the form

$$w \mapsto \frac{1}{f(1/w)} = \frac{w^n}{a_n w^n + a_{n-1} w^{n-1} + \ldots + 1},$$

which is a smooth function of w, provided that $|w|$ is sufficiently small (to avoid the denominator 0).

Problem. Find the degree of the resulting smooth map $\varphi : \mathbb{S}^2 \to \mathbb{S}^2$, (that is, $\mathbb{C}\mathrm{P}^1 \to \mathbb{C}\mathrm{P}^1$) defined by the polynomial f.

SOLUTION. The degree of the map φ equals the degree n of the polynomial f, provided that both spheres are oriented by the same orientation of the complex tangent line, "from 1 to i", at each point of the manifold $\mathbb{C}\mathrm{P}^1$.

Indeed, the preimages of a non critical value ω are the n roots of the corresponding equation of degree n, $f(z) = \omega$.

Lemma. *The signs of all these roots are positive.*

Proof. The non zero complex numbers form a connected set. So, the non-degenerate linear maps of a complex line form a connected set, and hence preserve the line orientation (it is the Italian principle discussed in Ch. 5). $\square$

Remark. This reasoning shows that the non-degenerate complex maps $\mathbb{C}^m \to \mathbb{C}^m$, considered as real linear maps $\mathbb{R}^{2m} \to \mathbb{R}^{2m}$, preserve the orientation. Hence, for the holomorphic maps all the points are positive, contributing $+1$ to the corresponding "degree" sum. This proves the

Proposition. *The topological degree of a polynomial of degree n is n.*

PROBLEM. Find the degree of the map $\varphi : \mathbb{S}^2 \to \mathbb{S}^2$ defined by the complex rational function $f(z) = p(z)/q(z)$, where p and q are complex polynomials with respective degrees m and n.

ANSWER. If the complex polynomials p and q are relatively prime, then $\deg \varphi = \max(m, n)$.

SOLUTION. To understand it, observe that for $f(z) = 1/z$, φ has degree 1 because the map $z \mapsto 1/z$ preserves the orientation (verify it!). For the rational function $f(z) = 1/q(z)$, φ has degree n because, as in the previous problem, the preimages of a non critical value ω are the n roots of the equation of degree n, $1 = \omega q(z)$. Finally, for $f(z) = p(z)/q(z)$ the preimages of a non critical value ω are the roots of the equation $p(z) - \omega q(z) = 0$, whose degree is $\max(m, n)$, that is, $\deg \varphi = \max(m, n)$.

The results of the preceding problems explain the motivation of the word "degree" for the integer that we have associated to the homotopy class of the n-spheroid of $\mathbb{S}^n$. It coincides with the algebraic degree of the case $n = 2$. For higher values of n our topological notion has no algebraic version of such simplicity, and we shall consider it independently of algebra.

10.2 Degree in differential equations theory

A very important application of the degree was suggested by Poincaré in his theory of differential equations. Consider in the affine space $\mathbb{R}^n$ a smooth vector field v, associating to any point $x \in \mathbb{R}^n$ a vector $v(x) \in T_x\mathbb{R}^n$. The usual coordinate system $x = (x_1, \ldots, x_n)$ provides n smooth functions for the components of the vector,

$$v(x) = (v_1(x_1, \ldots, x_n), \ldots, v_n(x_1, \ldots, x_n)).$$

To understand the construction of Poincaré, it is good to start with the case of plane vector fields, $n = 2$. Say, consider the following four vector fields, defining four systems of differential equations $\dot{\vec{x}} = \vec{v}(\vec{x})$:

$$\begin{cases} \dot{x} = x \\ \dot{y} = 2y \end{cases}, \quad \begin{cases} \dot{x} = y \\ \dot{y} = x \end{cases}, \quad \begin{cases} \dot{x} = y/2 \\ \dot{y} = -x/2 \end{cases}, \quad \begin{cases} \dot{x} = \ \ \varepsilon x + y/2 \\ \dot{y} = -x/2 + \varepsilon y \end{cases} (\varepsilon = 0.2)\,.$$

Problem. Draw the vector fields and sketch the trajectories, near the origin, of the points $(x(t), y(t))$ that verify the above differential equations

$$\begin{cases} \frac{dx}{dt} = P(x(t), y(t))\,, \\ \frac{dy}{dt} = Q(x(t), y(t))\ . \end{cases}$$

Answer. See Fig. 10.5.

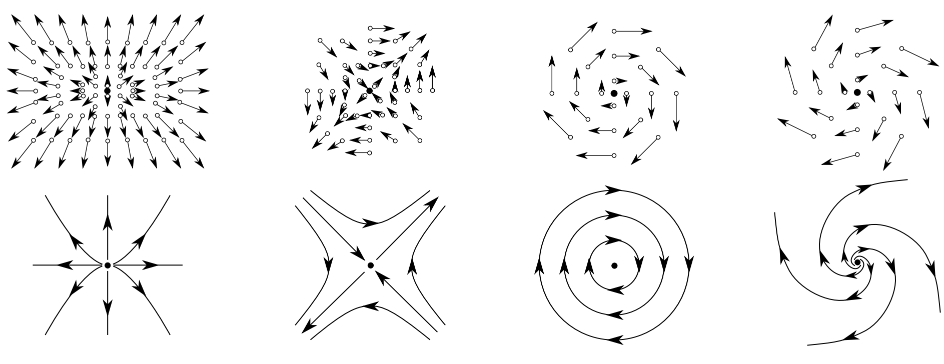


Figure 10.5: Singular points of plane vector fields: node, saddle, centre, focus.

Consider now a domain in $\mathbb{R}^n$, bounded by some closed smooth $n-1$ dimensional manifold γ (say, an embedded circle in the plane).

Singular Points. The *singular points* of the field are those where the field vanishes (that is, $v(x) = 0$ at a singular point x).

The singular points of a typical vector field on the plane are only nodes, saddles or foci (Fig. 10.5). These singular points are stable under small perturbations of the field.

Suppose that there is no singular point on the manifold γ (as it is the case for the circle $x^2 + y^2 = R^2$ in all examples of the preceding problem on plane vector fields) and suppose that the space $\mathbb{R}^n$ is oriented.

We orient the boundary $(n-1)$-manifold γ in the usual way: the frame of $\mathbb{R}^n$ which consists of the exterior normal vector to the domain bounded by γ followed by the orienting frame of γ, orient positively the space $\mathbb{R}^n$.

Index of a Vector Field. We associate an integer to γ and v. By definition, the *index of the field* v *along the hypersurface* γ is the degree of the map $\varphi : \gamma \to \mathbb{S}^{n-1}$, defined in the following way: For any $x \in \gamma$ we move the vector $v(x)$ from the space $T_x\mathbb{R}^n$ where it lives to the space $\mathbb{R}^n$ itself (translating it parallelly to the origin of $\mathbb{R}^n$), and then we normalise its length to become 1:

$$\varphi(x) = \frac{v(x)}{|v(x)|} \in \mathbb{S}^{n-1}, \quad \text{for } x \in \gamma.$$

Problem. Calculate the indices of the four vector fields of the preceding problem, along the curve $x^2 + y^2 = R^2$.

Remark. The orientations needed to define the degree are fixed here by the convention that γ is oriented as the boundary of the n-ball in the oriented space $\mathbb{R}^n$ and the unit sphere $\mathbb{S}^{n-1}$ is oriented as the boundary of the n-ball of shorter vectors.

These choices exclude the orientation of the ambient space $\mathbb{R}^n$ from the definition of the index of a field along γ. Indeed, a change of orientation of the ambient space would change both the orientation of the bounding hypersurface γ of $\mathbb{R}^n$ and of the sphere $\mathbb{S}^{n-1}$, preserving the value of the degree of the map φ defined by the vector field.

Answer. The indices for the fields of Fig. 10.5 are equal to $+1$, -1, $+1$ and $+1$, respectively.

To understand the notion of index, every one must compute it, independently of the teachers, for several examples of vector fields.

The index of a vector field has a remarkable additivity property: Decomposing a domain A (see Fig. 10.6) into two parts B and C, by a smooth section, we consider the indices of the field v along the three boundary surfaces ∂A, ∂B and ∂C.*

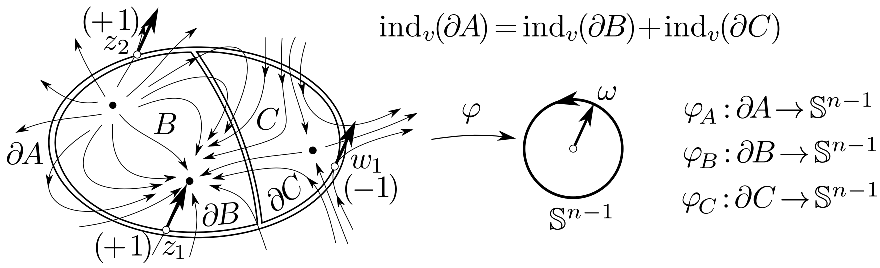


Figure 10.6: Index additivity theorem.

Theorem 2. *Any smooth vector field v satisfies*

$$\mathrm{ind}\,\partial A = \mathrm{ind}\,\partial B + \mathrm{ind}\,\partial C,$$

provided that the three boundary surfaces contain no singular point of v:

$$v(x) \neq 0 \quad \textit{for} \;\; x \in (\partial A \cup \partial B \cup \partial C).$$

Proof. Taking a non critical value $\omega \in \mathbb{S}^{n-1}$, we find its preimages $(z_1, \ldots, z_b)$ on ∂B and $(w_1, \ldots, w_c)$ on ∂C. The list $(z_1, \ldots, z_b, w_1, \ldots, w_c)$ contains the preimages of ω belonging to ∂A, since $\partial A \subset (\partial B \cup \partial C)$. The corresponding signs for ∂A are the same as for ∂B or for ∂C, since the boundaries are oriented in the same way.

The above list (possibly) contains some additional points on $\partial B \cap \partial C$, sent to ω by φ. Such points are counted twice in the list (as z-points and as w-points), but with opposite signs, since the exterior normals for B and for C are opposite at the points of the common boundary.

Thus, these common boundary points contribute nothing to the sum of all the signs of the points z and w of $\varphi^{-1}(\omega)$. Hence the theorem is proved. □

*In order to apply formally the previous definitions, the angles at the place of intersection of the section with the boundary of A should be smoothened, but in fact the index is well defined for the surfaces with angles too.

We can iterate the above theorem, decomposing the given domain into many small parts. The index is then the sum of the indices along the boundaries of the parts. We deduce an important result.

Lemma. *The index of a vector field along the boundary of a small domain, containing no singular points of that field, vanishes.*

Proof. The continuity of the field and the absence of singular points imply that the directions of the vectors of the field are almost parallel at all points of the small domain we are studying.

This proves the vanishing of the index on the boundary of this domain, since we may choose as a non critical value simply a point $\omega \in \mathbb{S}^{n-1}$, which is not the direction of the field at any of the points of our small domain. With such choice, the index is the sum of no summands, that is, it is zero. $\square$

In consequence, the index along the boundary of a neighbourhood of an isolated critical point does not depend on the choice of the neighbourhood: To arrive to a large neighbourhood, one should add several small domains which contribute nothing to the total sum.

Definition. The index of a vector field along the spherical boundary of a small ball centred at an isolated singular point of that field, is called *the index of the field at this singular point.*

PROBLEM. Calculate the indices of the fields of Fig. 10.5 at the origin.

ANSWER. $+1$, -1, $+1$, $+1$.

PROBLEM. Find smooth vector fields in the plane of indices -2 and $+2$ at the origin.

ANSWER. See Figure 10.7 for which the corresponding differential equations are $\dot{z} = z^2$ and $\dot{z} = \overline{z}^2$ (in the plane $\mathbb{R}^2$ of the variable $z = x + iy$).

PROBLEM. Find the indices of the critical points of the field of equation $\dot{z} = 1$, prolonged from $\mathbb{C}$ to $\mathbb{C}\mathrm{P}^1$ by continuity.

ANSWER. This field is $\dot{w} = -w^2$ for the local coordinate $w = 1/z$ in $\mathbb{C}\mathrm{P}^1$. Hence the solution is provided by the preceding problem (and by Fig. 10.7).

PROBLEM. Given a non-degenerate linear operator $A : \mathbb{R}^n \to \mathbb{R}^n$, find the index of the linear field of the differential equation $\dot{x} = Ax$.

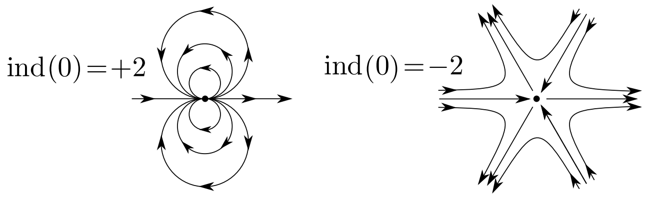


Figure 10.7: Lines of the fields $v(z) = z^2$ and $v(z) = \overline{z}^2$.

Answer. Let k be the number of negative eigenvalues of the operator A whose eigenvalues are real, counting the multiple eigenvalues with their multiplicities. Then $\operatorname{ind} v$ equals $(-1)^k$ at the point 0.

Hint. Consider the diagonal system $\dot{x}_j = \lambda x_j \quad (j = 1, \ldots, n)$, and reduce the given field to this one by a homotopy.

Remark that a pair of complex conjugate eigenvalues would contribute 1 to the index, like for $(A) = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. For instance, one could homotopically deform the pair of complex conjugate eigenvalues to the pair $(1, 1)$, and the index would remain constant along that homotopy.

Problem. Prove that the index of an isolated singular point of a vector field does not depend on the coordinate system implicitly used to define this index.

Solution. Connecting two coordinate systems (orienting the space in the same way) by a continuous one-parameter family of coordinate systems, we obtain on the boundary sphere of a small neighbourhood of the origin a vector field with continuously changing components, depending on the parameter of the family.

The resulting homotopy of maps to the unit sphere proves the equality of the indices of the field at the singular point.

The independence remains true for the orientation reversing coordinate changes, since the index is obviously preserved by the standard reflection which changes the sign of just one coordinate.

10.2.1 Degree of the Gauss Map

One more example of the usefulness of the degree of a map from an oriented m-manifold M to the oriented sphere $\mathbb{S}^m$ is given when M^m is the smooth boundary of a domain B^{m+1} in Euclidean space $\mathbb{R}^{m+1}$.

We equip the hypersurface M with the exterior normal unit vector $\nu(x)$ at each point x. Translating this vector parallelly to the origin, we get a vector $\Gamma(x)$ of the unit sphere $\mathbb{S}^m \subset \mathbb{R}^{m+1}$ centred at the origin. The map Γ is called the *Gauss map* of M to $\mathbb{S}^m$.

Problem. Calculate the degree of the Gauss map of the surfaces shown in Fig. 10.8.

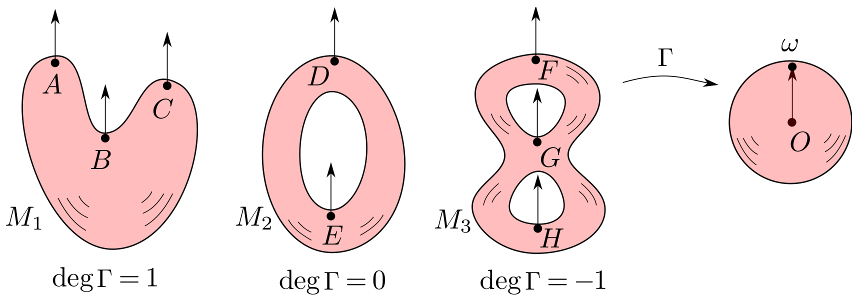


Figure 10.8: The preimages of the point $\omega \in \mathbb{S}^2$ for the Gauss maps Γ of the surfaces M_1, M_2, M_3.

Remark. We suppose here that M is oriented as the boundary of the domain B^n and $\mathbb{S}^m$ – as the boundary of the set of points inside the ball bounded by this sphere. The resulting degree does not depend on the orientation of the space $\mathbb{R}^n$.

Solution. Take a non critical value $\omega \in \mathbb{S}^2$ (the “North” pole in Fig. 10.8). It is easy to count its preimages – they are those points of the surface where the exterior normal is vertical and directed up. To compute the sign of such a point, we consider locally the surface as the graph a smooth function $z = f(x, y)$ of two variables on the horizontal plane such that

$$M = \{(x, y; z) : z = f(x, y)\}, \quad f(0,0) = 0, \quad df(0,0) = 0.$$

The sign of our point is the sign of the Hessian of the function $-f$, that is, the sign of Gaussian curvature, which is the determinant of the second fundamental form (see p. 314 and Fig. 9.2). It remains to observe that, at the points with vertical upward normal, the Gaussian curvature signs are:

point	A	B	C	D	E	F	G	H
sign	+	−	+	+	−	+	−	−

.

Hence, we conclude that the degree of the Gauss map $\Gamma : M \to \mathbb{S}^2$ is, respectively, $+1$ for the spherical surface M_1, 0 for the torus surface M_2 and -1 for the surface M_3 (of genus 2).

Remark. Deforming continuously the surface, we obtain a homotopy of the Gauss map, and hence the degree does not change. So, it is equal to 1 for every surface bounding a ball, like an ellipsoid and so on; it is equal to 0 for any toric surface; and to -1 for any surface of genus 2.

Exercise. Check that the Gauss map of a surface of genus g has degree $1-g$.

10.2.2 Euler Characteristic

We start by solving the following problem

Problem. Calculate the sum of the indices of the singular points of any vector field on the 2-sphere $\mathbb{S}^2$ (whose singular points are all isolated).

Solution. Consider a circular neighbourhood U of any non singular point of the vector field. The complement $\widetilde{U}$ of U in $\mathbb{S}^2$ is diffeomorphic to a disc in the plane $\mathbb{R}^2$ – Fig. 10.9.

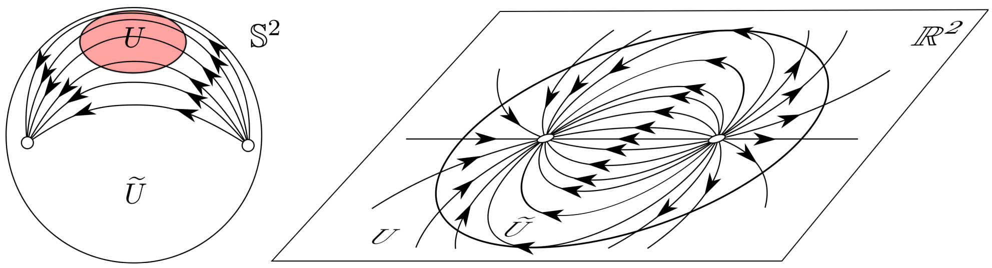


Figure 10.9: Proof of the independence of the sum of the indices of the singular points of a vector-field on $\mathbb{S}^2$.

The index of the restriction of the vector field to the boundary curve of $\widetilde{U}$ can be computed as the index of the field on the disc U along its boundary circle (taken with the sign, depending on the orientation of the boundary, which is opposite for the boundaries of the domains U and $\widetilde{U}$ of the same oriented 2-sphere $\mathbb{S}^2$).

So, the sum of the indices of the singular points of the field on $\mathbb{S}^2$ is equal to the index of its restriction to the boundary curve, oriented as $\partial\widetilde{U} = -\partial U$.

According to Fig. 10.7 (left), p. 370, the index of the restriction of this almost parallel field in U to the boundary curve, considered in the plane $\mathbb{R}^2$ where lives the diffeomorphic representative of the domain $\widetilde{U}$, equals $+2$ (for the orientation of $\partial\widetilde{U}$).

Thus, the sum of indices over all singular points of the original field on the sphere $\mathbb{S}^2$ is also $+2$, independently of the continuation of this field to the remaining part $\widetilde{U}$ of the sphere $\mathbb{S}^2$.

Problem. Find the sum of the indices of the singular points of a smooth vector field on the sphere $\mathbb{S}^n$ (supposing the singular points to be isolated).

Answer. For n even the sum is always $+2$, like for $\mathbb{S}^2$. For n odd the sum is always 0, like for $\mathbb{S}^1$.

Remark. It is easy to construct a smooth vector field without singular points on any odd-dimensional sphere $\mathbb{S}^{2k-1}$. Namely, at each point z of the unit sphere in the complex vector space $\mathbb{C}^k$, we associate the vector $v(z) = iz$ (which is orthogonal to z and hence is tangent to the sphere).

Problem. Find the sum of the indices of a smooth vector field on a torus $\mathbb{T}^2$, whose singular points are isolated.

Answer. This sum is zero for any field. This follows from the vanishing of the index for the "parallel vector field" on the covering plane $\mathbb{R}^2$, which has no singular points. But to use this particular field, we are applying the solution of the following problem.

Problem. Prove that *the sum of the indices of all the singular points of a smooth vector field on a smooth orientable compact manifold is independent of that field* (which is supposed to have only a finite number of singular points).

Solution. To prove it, we represent the whole manifold as the union of small squares (cubes) whose boundaries contain no singular point of the vector field. The index of the field along the boundary of such small square (cube) remains constant for a small deformation of the field.

Since the sum of the indices of all the singular points is equal to the sum of the indices along the boundary squares (cubes), it takes the same value for the initial field and for any sufficiently close vector field (even if the number of singular points of those two fields are different).

The main point is that any two vector fields can be connected by a finite sequence of such small deformations. Now, since every small deformation preserves the sum of the indices of the singular points, this sum takes equal values at the initial and at the final vector field.

Problem. Find the sum of the indices of any smooth vector field on a surface of genus g, supposing the field has a finite number of singular points.

Solution. The sum is $2 - 2g$. It suffices to prove it for a particular field. Consider the linear function $y : M^2 \to \mathbb{R}$, defined by the orthogonal projection of the Euclidean space $\mathbb{R}^3$ to a generic line, called "vertical" (as that of Fig. 10.8).

Consider the gradient vector field of this function on the Riemannian manifold M^2 (we are considering the Riemannian metric on $M \subset \mathbb{R}^3$ induced by the Euclidean structure on $\mathbb{R}^3$). Its singular points are the preimages of the two vertical unit vectors ω and $-\omega$, for the Gauss map $M \to \mathbb{S}^2$.

The index of the gradient field at such a singular point is equal to the sign of this preimage point of the Gauss map, assuming the non-degeneracy of the second quadratic form (which is the case for a generic choice of the "vertical" direction). The sum of the indices of the singular points for this particular field, which depends on the embedding of the surface to $\mathbb{R}^3$ (which we can embed as we wish) is equal to the doubled degree of the Gauss map (taking into account both the contributions of the preimages of the North pole ω and of the South pole $-\omega$).

Now, taking the standard embedding of the surface of genus g (shown in Fig. 10.8 for values of the genus $g = 0, 1, 2$), we prove the required formula for this field v

$$\sum_{x \in M : v(x)=0} (\operatorname{ind} v \text{ at } x) = 2 \deg(\Gamma : M \to \mathbb{S}^2) = 2 - 2g. \tag{2}$$

Euler Characteristic. The sum of indices of the singular points of a vector field on a compact manifold M is called its *Euler Characteristic*, $\chi(M)$.

As we have just proved, it is independent of the field, and hence, to compute $\chi(M)$, it suffices to study just one specific field.

Problem. Find the Euler characteristic of the spheres, of the real and complex projective spaces and of the groups $\mathrm{SO}(n)$ and $\mathrm{U}(n)$.

Answer.

$$\chi(\mathbb{S}^{2k}) = 2, \quad \chi(\mathbb{S}^{2k-1}) = 0,$$
$$\chi(\mathbb{R}\mathrm{P}^{2k}) = 1, \quad \chi(\mathbb{R}\mathrm{P}^{2k-1}) = 0,$$
$$\chi(\mathbb{C}\mathrm{P}^n) = n+1, \quad \chi(\mathrm{SO}(n)) = 0, \quad \chi(\mathrm{U}(n)) = 0.$$

Theorem 3. *The Euler characteristic of every odd dimensional compact manifold M is equal to zero*: $\chi(M) = 0$ *for odd* $n = \dim M$.

Proof. Consider a generic vector field v on M, i.e., with isolated non degenerate singular points (for example, the gradient vector field of a generic smooth function $f : M \to \mathbb{R}$). The index of v at a singular point p is $\mathrm{ind}_p v = (-1)^\nu$, where ν is the number of negative eigenvalues of the linearised field of v at p, $\dot{x} = Ax$ (see p.370). Thus the index of the reversed field $-v$ at the same point p is equal to $(-1)^{n-\nu} = -(-1)^\nu = -\mathrm{ind}_p v$ because n is odd. Hence, the sums of indices (at the singular points) for v and for $-v$ have opposite sign. Since both sums are equal to $\chi(M)$, we have $\chi(M) = 0$. □

Theorem 4. *For any compact smooth group G we have* $\chi(G) = 0$.

Proof. To construct a vector field on a compact smooth group manifold G, it suffices to take any non zero tangent vector v_0 at the unity point e of G, and to translate it to any point g of G by the left translation diffeomorphism on G defined by $L_g : G \to G$, $L_g(h) = gh$. For any $h \in G$, $v(g) = L_{g*}v_0$.

The resulting field $v(g) \in T_gG$ (invariant by left translations) is evidently nowhere vanishing. □

Corollary. *No complex projective space is a smooth group manifold. No even-dimensional sphere is diffeomorphic to a smooth group.*

We have already seen that the spheres $\mathbb{S}^1$ and $\mathbb{S}^3$ are smooth group manifolds (formed, respectively, by the norm 1 complex numbers and the norm 1 quaternions).

The sphere $\mathbb{S}^5$ is not diffeomorphic to a smooth group. We shall not prove this assertion, but it follows from the non parallelisability of $\mathbb{S}^5$ (there is no smooth field of orthogonal 5-frames in its tangent spaces), and from the

Proposition. *Every smooth group manifold is parallelisable.*

Proof. One translates any initial frame f_0 from the tangent space at e to the other tangent spaces by the left translations. □

Theorem. *The only parallelisable spheres are* $\mathbb{S}^1$*,* $\mathbb{S}^3$ *and* $\mathbb{S}^7$*.*

We shall not prove in this book this celebrated theorem, which is one of the best results of the XXth century mathematics.

10.3 Gauss-Bonnet Theorem

Problem. Calculate the integral of the Gaussian curvature of an ellipsoid M in Euclidean space $\mathbb{R}^3$.

Solution. Use the image point $\Gamma(x) = \omega \in \mathbb{S}^2$ of the Gauss map as a local coordinate of the point $x \in M$, and denote by ds and $d\omega$ the area elements of the surfaces M and $\mathbb{S}^2$. Thus we get the identity

$$\oiint_M K(x)ds = \oiint_M d\Gamma(x) = \oiint_{\mathbb{S}^2} \sum(\omega)d\omega,$$

where

$$\sum(\omega) = \sum_{\Gamma(x)=\omega} \operatorname{sign}\det\left(\frac{\partial\Gamma}{\partial x}\right), \quad \text{since} \;\; d\omega = \det\left(\frac{\partial\Gamma}{\partial x}\right)dx.$$

Taking into account the formula for the degree,

$$\sum_{\Gamma(x)=\omega} \operatorname{sign}\det\left(\frac{\partial\Gamma}{\partial x}\right) = \deg\Gamma,$$

which is true for all non critical values ω (and hence almost everywhere), we get

$$\oiint_M K(x)ds = \oiint_{\mathbb{S}^2}(\deg\Gamma)d\omega = 4\pi\deg\Gamma = 4\pi.$$

Thus the integral of the Gaussian curvature of an ellipsoid is equal to 4π.

Since the proof is based on the formula for the degree, the result is valid for any surface homotopic to the sphere.

By this reasoning and formula (2) (of p.374) we obtain the

Gauss-Bonnet Theorem. *For the surfaces of higher genus* g*, we have*

$$\oiint_M K(x)ds = 4\pi\cdot\deg\Gamma = 4\pi\cdot\frac{1}{2}\chi(M) = 2\pi(2-2g)\,.$$

Hypersurfaces of dimension m. The same reasoning provides a similar formula

$$\int \cdots \int_M K(x)ds = a_m \cdot \deg \left(\Gamma : M^m \to \mathbb{S}^m \right),$$

where a_m is the m-dimensional area of the hypersphere of radius 1 in Euclidean space $\mathbb{R}^{m+1}$.

Notice that for even m the number $a_m \cdot \deg(\Gamma : M^m \to \mathbb{S}^m)$ is a topological invariant which is independent of the immersion (or of the embedding) of the hypersurface M^m in Euclidean space $\mathbb{R}^{m+1}$, while the Gaussian curvature depends on that immersion (embedding); however, if m is odd, $\deg \Gamma$ can be arbitrary (see the example for $m = 1$ below). Thus if m is even, then $\chi(M^m) = \deg \Gamma \cdot \chi(\mathbb{S}^m) = 2 \deg \Gamma$, that is, $\deg \Gamma = \frac{1}{2}\chi(M^m)$ and

$$\int \cdots \int_M K(x)\, ds = \frac{1}{2}\, \chi(M^m) \cdot a_m \qquad \text{(for even } m\text{)}.$$

Problem. Find the integral of the Gaussian curvature along the closed curve "∞" immersed in Euclidean plane $\mathbb{R}^2$.

Solution. In this example the length of the circle of radius one is $a_1 = 2\pi$, and hence

$$\oint_{\text{“}\infty\text{”}} K ds = 2\pi \deg(\text{“}\infty\text{”} \to \mathbb{S}^1).$$

Since the degree of the Gauss map is equal to 0 for this immersion, we get the nontrivial identity

$$\oint_{\text{“}\infty\text{”}} K ds = 0,$$

whatever asymmetric realisations of the curve "∞" are studied.

For other immersions of the circle $\{t \pmod{2\pi}\}$ into the plane $\mathbb{R}^2 = \{(x, y)\}$, like $\{x = \cos t,\ y = \sin t\}$ and $\{x = \cos 2t,\ y = \sin 2t\}$, the integral of the curvatures take different values (2π and 4π in these two examples).

10.3.1 Euler-Poincaré Formula for Polyhedra

Problem. Let P be a convex polyhedron in Euclidean space $\mathbb{R}^3$ with v vertices, e edges and f faces. Calculate the number $v - e + f$.

Solution. Consider an interior point O in the domain bounded by the surface P, and let $d : P \to \mathbb{R}$ be the function whose value at a point $x \in P$ is the distance from O to x (see Fig. 10.10),

$$d(x) = |x - O|.$$

Smoothening the surface and the function d, we consider the vector field $\operatorname{grad} \widetilde{d}$ on the smooth surface $\widetilde{P}$ diffeomorphic to the sphere $\mathbb{S}^2$ (shown in Fig. 10.10 for the cube surface P).

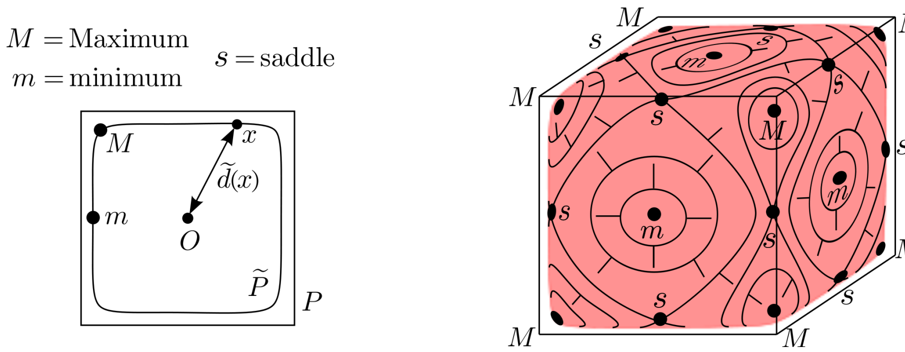


Figure 10.10: Smoothened distance function $\widetilde{d}$ on the smoothened cube surface.

The sum of the indices of the critical points of this function is $\chi(\mathbb{S}^2) = 2$, as we have proved above.

But any suitable smoothened function $\widetilde{d}$ has a local maximum near each vertex of P, a local minimum at one point of each face and a saddle point at one point of each edge. Thus,

$$\begin{aligned} v &= \text{number of local maxima points (8 for the cube)} \\ e &= \text{number of saddle points (12 for the cube)} \\ f &= \text{number of local minima points (6 for the cube)} \end{aligned}$$

and these are all the critical points of $\widetilde{d}$, that is, all the singular points of its gradient field. Thus, for any convex polyhedron we get the astonishing identity

$$v - e + f = 2.$$

This formula (proved by Pascal and by Descartes, who triangulated the faces) is called *Euler's formula*. The alternating sum $v - e + f$ is called the *Euler characteristic* of the polyhedron P.

For the polyhedral surfaces P topologically different, the formula says

$$v(P) - e(P) + f(P) = \chi(P).$$

Its independence of the polyhedral representation of the surface P is an unevident theorem (proved above for the smoothable surfaces).

Its Poincaré version holds for the higher dimensional polyhedra:

$$\chi(P^n) = \sum_{j=0}^{n} (-1)^j a_j(P),$$

where P^n is a polyhedral surface with $a_j(P)$ convex faces of dimension j: a_0 vertices, a_1 edges,…, a_n n-dimensional convex faces.

The celebrated elementary geometric statement is the astonishing fact that the sum in the right hand side of the above *Euler-Poincaré formula* is independent of the special choice of the polyhedron P^n into the class of the polyhedra homeomorphic to it. Say, for the cube, the tetrahedron and the dodecahedron this formula provides equal alternating sums

$$(8 - 12 + 6) \;=\; (4 - 6 + 4) \;=\; (20 - 30 + 12) \;=\; 2 \;\; (= \chi(\mathbb{S}^2)).$$

We shall discuss later the topologic characteristics of the polyhedra P^n, reflected by the value of its Euler characteristic $\chi(P^n)$.

Exercise. Compute the Euler characteristic of the n-dimensional simplex and of the n-dimensional cube in Euclidean space $\mathbb{R}^n$:

$$\textit{Simplex}: \quad \{x \in \mathbb{R}^n : x_j \geqslant 0 \quad (\text{for } j = 1, \ldots, n) \text{ and } \textstyle\sum_{j=1}^n x_j \leqslant 1\},$$

$$\textit{Cube}: \quad \{x \in \mathbb{R}^n : 0 \leqslant x_j \leqslant 1 \quad (\text{for } j = 1, \ldots, n)\}.$$

10.4 Index of intersection

The index of a singular point of a vector field can be interpreted by the following slightly more general construction.

Consider two smooth oriented submanifolds, X^k and Y^ℓ, in an oriented manifold Z^n whose dimension equals the sum of the dimensions of those submanifolds, $n = k + \ell$ – Fig. 10.11.

Suppose that the submanifolds are not tangent at their intersection points.

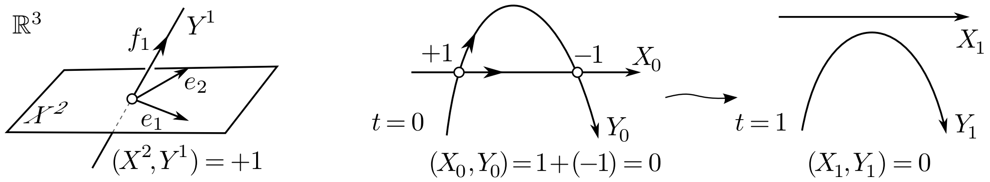


Figure 10.11: Intersection index definition and invariance under deformations.

The union of the orienting frames $(e_1, \dots, e_k)$ of X^k and $(f_1, \dots, f_\ell)$ of Y^ℓ at their intersection points, taken in this order $(e_1, \dots, e_k, f_1, \dots, f_\ell)$, form a frame of the ambient manifold Z^n.

Definition. The *index of intersection of* X^k *and* Y^ℓ *at a point of intersection* is $+1$ if the resulting n-frame orients the manifold Z^n positively, and -1 if it orients Z^n negatively. The *index of intersection* of two oriented submanifolds X^k and Y^ℓ of an oriented manifold Z^n is the sum of the intersection indices taken over all their intersection points.

We are supposing that the submanifolds are closed, of complementary dimensions ($n = k+\ell$) and they intersect transversely, that is, at any of their intersection points their tangent spaces have no nonzero common vector.

Theorem 5. *The intersection index* (X^k, Y^ℓ) *in* Z *remains constant if the submanifolds* X *and* Y *experience homotopies.*

The proof is quite similar to the proof of the degree invariance (see Fig. 10.3 on page 361), and we leave the details to the reader.

In the case of an isolated tangency point (point of non transverse intersection) of X and Y, we may define the intersection index at that point as the sum of the intersection indices at the transverse intersection points of some deformed manifolds $\widetilde{X}$ and $\widetilde{Y}$, in some neighbourhood of the initial point with “bad” –non transverse– intersection.

This integer does not depend on the choice of the deformed manifolds $\widetilde{X}$ and $\widetilde{Y}$, according to the preceding theorem, provided that the deformation is sufficiently small.

Problem. Calculate the intersection index of the surfaces given by the equations $w = 0$ and $w = z^2$, at the origin of the complex 2-plane with coordinates $z \in \mathbb{C}$ and $w \in \mathbb{C}$.

Answer. For the n-th order tangency ($w = 0$ and $w = z^n + \ldots$) the intersection index (at 0) equals n.

Problem. calculate the intersection index of the surfaces $z = t = 0$ and $z = x^2 + y^2$, $t = 2xy$ in the space $\mathbb{R}^4$ with coordinates (x, y, z, t).

Answer. The intersection index is zero.

Theorem 6. *The index of a singular point of a smooth vector field v in $\mathbb{R}^n$ is equal to the intersection index of its graph $\{x, y = v(x)\}$ with the "horizontal plane" $y = 0$ in the vector space $\mathbb{R}^{2n} = \{(x, y)\}$.*

Proof. To simplify the notation, we suppose the intersection is $x = 0$. The tangent n-plane to the graph $\{x, y = v(x)\}$ is the graph $\{x, y = Ax\}$ of the linearised vector function v.

Thus, in the non-degenerate case ($\det A \neq 0$), we get just the same definition: both the index of intersection and the vector field index are equal to the sign of the determinant. $\square$

Remark. The orientation of the space $\mathbb{R}^n$ has no importance in this case, since a change of the orientation of $\mathbb{R}^n = \{x\}$ would change also the orientation of its tangent space $\mathbb{R}^n = \{y\}$, and hence the orientation of the "graph space" $\mathbb{R}^{2n}$, as well as both the vector field index and the intersection index remain unchanged.

This remark explains why the indices of the singular points of a vector field are well defined on non oriented and on non orientable manifolds.

In the degenerate case (of an isolated singular point) the index of the vector field still coincides with the index of intersection, but we leave the details to the reader (in both cases, to avoid the tangency, one should perturb v by a deformation).

10.4.1 Self-intersection Number

Even in the case of the coincidence of the submanifolds X and Y in Z, one defines their intersection number, (X, X), called the *self-intersection number*. To calculate it, one replaces $Y = X$ by its perturbed version $\widetilde{Y}$. The result is independent of the particular choice of the perturbation.

Problem. Find the self-intersection number of a straight line $X = \mathbb{C}\mathrm{P}^1$ in the complex projective plane $\mathbb{C}\mathrm{P}^2$.

Answer. $(X, X) = 1$: Take for $\widetilde{Y}$ a projective straight line slightly different from X.

Problem. Find the self-intersection number of the elliptic curve given by

$$y^2 = x^3 - x,$$

in the complex projective plane (with affine coordinates x and y) – Fig. 10.12.

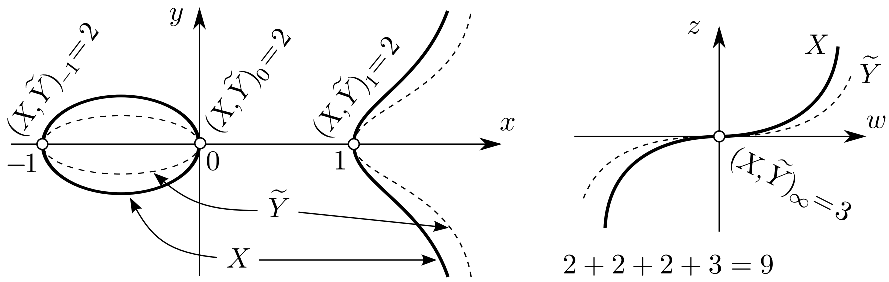


Figure 10.12: Selfintersection number of an elliptic curve in $\mathbb{C}\mathrm{P}^2$.

Solution. The self-intersection number is 9. One can deform the curve X to become $\widetilde{Y}\ :\ (1-\varepsilon)y^2 = x^3 - x$.

The four intersection points of the elliptic curves X and $\widetilde{Y}$ are

$$(x,\ y) = (-1,\ 0),\ (0,\ 0),\ (1,\ 0) \text{ and } (\infty,\ \infty).$$

At each of the first three points, the tangency is quadratic, whence the intersection indices at these three points equal 2.

Near the last intersection point the local coordinates of $\mathbb{C}\mathrm{P}^2$ are $z = 1/y$ and $w = x/y$. The equation of the elliptic curves X and $\widetilde{Y}$ in these coordinates are $z = w^3 - wz^2$ and $(1-\varepsilon)z = w^3 - wz^2$. So, the Taylor series are $z = w^3 + o(|w|^3)$, $z = (1/(1+\varepsilon))w^3 + o(|w|^3)$, and hence we have the cubic tangency at the point $z = w = 0$. Thus the intersection index at this point equals 3.

In conclusion, the self-intersection number of the elliptic curve X equals $2 + 2 + 2 + 3 = 9$.

Problem. Find the self-intersection number of a generic algebraic curve of degree n in $\mathbb{C}\mathrm{P}^2$.

Solution. Consider the completely degenerated curve which consist of n straight lines. To calculate the self-intersection number, we deform these lines, and we obtain the number of intersections n^2 with the deformed curve.

The smoothening of the singular points of the degenerated curve are local and occur far from these n^2 points of intersection with the deformed curve. The intersection number remains the same for the close generic curves of degree n. Hence, it is always n^2, according to the general “Italian principle” (Ch. 5, p. 145). For instance, we obtain the same self-intersection number $9 = 3^2$ for the cubic elliptic curve of the preceding problem.

Problem. Find the self-intersection number of the sphere $\mathbb{S}^2$, embedded into its own tangent bundle $T\mathbb{S}^2$ as the zero section.

Solution. As a perturbation we can take a section defined by a small generic vector field. The intersection points of this section with the zero section are the singular points of the field. The intersection numbers at these points are the indices of these singular points of the vector field. In consequence, we get the

Theorem. *The self-intersection number of the zero section of the tangent bundle TM is equal to the Euler characteristic $\chi(M)$ of the manifold M.*

So, to solve the above problem, we use that $\chi(\mathbb{S}^2)$ is equal to 2.

Remark. The oriented ambient manifold Z^{k+l}, which contains the two intersecting submanifolds X^k and Y^l, does not need to be compact (and is not compact in the preceding problem, $Z = TM$).

Towards Intersection Ring and Homology The above condition $\dim Z = \dim X + \dim Y$ can be weakened also, to $\dim X + \dim Y \geqslant \dim Z$. In this case the transverse intersection manifold $X \cap Y$ is generically of positive dimension $s = \dim X \cap Y = \dim X + \dim Y - \dim Z$.

There is a generalised version of the index of intersection which provides an orientation of the s-dimensional intersection variety $X \cap Y$. This construction is the base of the Poincaré *intersection ring* of a smooth manifold.

In that theory, the theorem of invariance under deformations holds (in a suitable form for these intersections) but Poincaré replaced it by a more powerful property.

Namely, instead of the homotopy classes, one considers some larger equivalence classes called homology classes.

Two closed oriented curves in a manifold are said to be *homologous* if there is an oriented surface whose boundary consists of these two curves with alternating orientations (Fig. 10.13). In the case when one of these two curves void, the other curve is said to be homologous to zero, as it happens to the curve γ on the surface of genus 2 of Fig. 10.13.

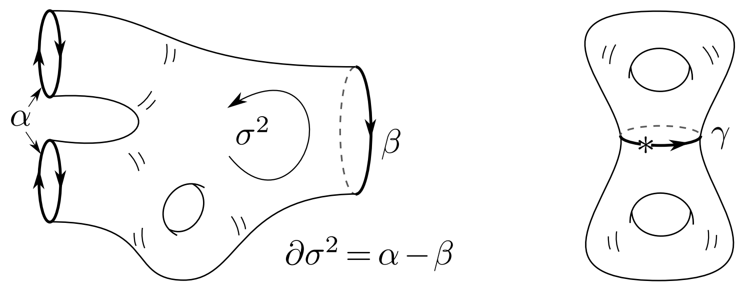


Figure 10.13: Left: cobordism and homology σ^2 of the curves α and β. Right: a curve γ homologous to 0 on a surface of genus $g = 2$, but representing a non-trivial element of its fundamental group.

Similar homology classes of any dimension k are defined as the equivalence classes of the boundary parts of a $(k+1)$-dimensional object. If the considered objects are smooth manifolds one works with the so-called *cobordism theory*.

However, a century of long investigations has shown that it is also very important to consider more general (non smooth) objects, called *chains* (as we have done in Chapter 6). For instance the boundary of a polyhedron is not a smooth manifold, but a 2-chain; the three sides of a triangle form a 1-chain, and so on. The main property of a chain is the possibility to integrate along it. The resulting theory of the homology classes is simply called *homology theory*, and the elements of the Poincaré intersection ring are such classes. The next chapter is devoted to homology theory.

10.5 Linking number theory

One more (extremely useful) example of application of the degrees of the maps is the so-called linking theory.

We start from the simple example of two smooth closed disjoint curves in oriented Euclidean space $\mathbb{R}^3$ (Fig. 10.14).

The *linking number* $L(\alpha, \beta)$ of two such curves $\alpha : \mathbb{S}^1 \to \mathbb{R}^3$, $\beta : \mathbb{S}^1 \to \mathbb{R}^3$ is defined by the following construction.

One takes an oriented surface (or even an oriented 2-chain) σ^2 whose (oriented) boundary is the first curve, $\partial\sigma^2 = \alpha$.

Then one takes the intersection index of the second curve with the surface σ^2. This integer is called the *linking number* of the oriented curves α and β in the oriented 3-space,

$$L(\alpha, \beta) = (\sigma^2, \beta), \quad \text{where } \partial\sigma^2 = \alpha.$$

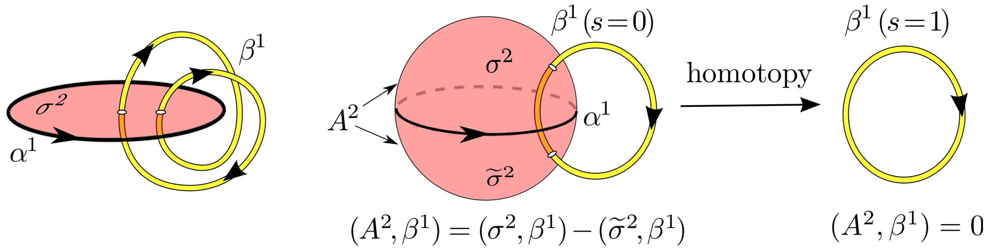


Figure 10.14: Left – $L(\alpha^1, \beta^1) = 2$. Right – The liking number is independent of the choice of the surface σ^2 with boundary α^1.

Theorem 7. *The linking number does not depend on the choice of the oriented surface* (*or* 2-*chain*) σ^2.

Proof. Let $\widetilde{\sigma}^2$ be a different oriented surface (or 2-chain) with equal boundary, $\partial\widetilde{\sigma}^2 = \partial\sigma^2 = \alpha$. We construct from these two surfaces (or chains) a closed surface (that is, compact without boundary):

$$A = \sigma^2 - \widetilde{\sigma}^2, \quad \partial A = \alpha - \alpha = 0.$$

This closed surface is the union of the two parts σ^2 and $\widetilde{\sigma}^2$, but the orientation of the second part is reversed (to eliminate the boundary), as it happens for the two equally oriented hemispheres: The boundary of the hemispheres is the same equator curve, but its orientation as the boundary of the Northern hemisphere is different from its orientation as the boundary of the South hemisphere, provided that both hemispheres are oriented in the same way, by the orientation of the whole sphere.

In our case, the two surfaces have equally oriented boundary curve α, and hence to connect them in one oriented surface without boundary one should reverse one of the orientations.

The intersection index of the resulting closed surface $A = \sigma^2 \cup \widetilde{\sigma}^2$ with the closed curve β in $\mathbb{R}^3$ is equal to 0. To prove it, we make a homotopy by translating the closed curve β far from the surface A – Fig. 10.14.

At the end of the homotopy the intersection index will be zero, since there will be no intersection point. The intersection index ought to be 0 also initially, since it is invariant under homotopy. Thus,

$$(A, \beta) = 0, \quad \text{that is,} \quad (\sigma^2, \beta) - (\widetilde{\sigma}^2, \beta) = 0,$$

which proves that the linking number is independent of the choice of the particular surface σ^2. $\square$

The linking number of closed curves was extremely useful in the magnetic field studies, providing important formulae in the theory of the electromagnetic field.

An explicit integral formula for the linking number was proved by Gauss:

$$L(\alpha,\beta)=\frac{1}{4\pi}\oiint\frac{[\frac{d\alpha}{ds},\frac{d\beta}{dt},F(s,t)]}{\mid F(s,t)\mid^3}dsdt\,, \tag{3}$$

where $F(s,t)=\alpha(s)-\beta(t)$. Here s and t are real parameters of the first and second immersed disjoint curves

$$\alpha(s)\in\mathbb{R}^3,\quad \beta(t)\in\mathbb{R}^3,\quad \alpha(s)\neq\beta(t).$$

The triple product symbol $[\xi,\eta,\zeta]$ means the oriented volume of the parallelepiped with edges ξ, η and ζ, equal to the determinant of their components.

The Gauss formula is based in the following theorem.

Theorem 8. *The linking number of two oriented disjoint curves α and β in oriented Euclidean space $\mathbb{R}^3$ coincides with the degree of the following "Gauss map" of the oriented 2-torus $\mathbb{T}^2=\mathbb{S}^1_1\times\mathbb{S}^1_2$, parametrised by s and t, to the oriented unit sphere of $\mathbb{R}^3$:*

$$g(s,t)=\frac{F(s,t)}{\mid F(s,t)\mid}.$$

Proof. It is evidently true for two curves separated by a large distance, since both the degree and the linking number vanish.

Deforming one of the curves α and β, one may continuously attain any configuration. However, in this homotopy, one has to pass by cases of intersecting curves.

One easily computes each contribution of the local deformation when the homotopy pass by a simple crossing of the curves, say, remaining parallel to itself during all this deformation.

To check the contributions of the crossing to the linking number and to the degree of the "Gauss map" is an explicit elementary geometric problem. Both contributions are either $+1$ or -1, but with equal signs in both cases.

Applying this reasoning a finite number of times (equal to the number of crossings needed to unlink the two curves), we conclude that the linking number is always equal to the degree of the Gauss map.

Now, to prove the formula for the degree, consider the conic sector based in a small base domain of oriented area $d\omega$ at the point $\omega = g(s,t) \in \mathbb{S}^2$ – Fig. 10.15 (the sector is formed by the segments connecting the centre O of the sphere to this small base domain).

The oriented volume of such an infinitesimal cone based on the two tangent vectors of the sphere $(\partial g/\partial s)ds$ and $(\partial g/\partial t)dt$ at the point $g(s,t)$, is equal to one third of the triple product,

$$dV = \frac{1}{3}\left[\frac{\partial g}{\partial s}ds,\ \frac{\partial g}{\partial t}dt,\ g\right].$$

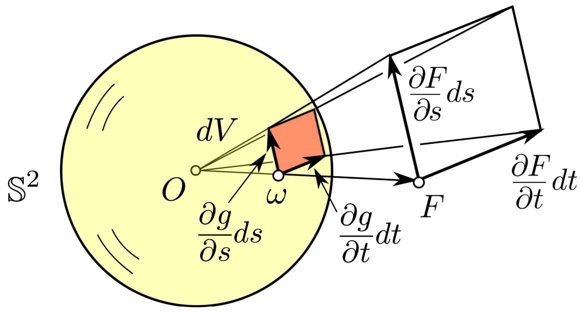


Figure 10.15: Proof of the Gauss formula for the linking number of two closed curves.

We calculate the differentials of g in terms of those of the vector-function $F = g \mid F \mid$, which have the form (Fig. 10.15):

$$\frac{\partial F}{\partial s} =\mid F \mid \frac{\partial g}{\partial s} + aF, \qquad \frac{\partial F}{\partial t} =\mid F \mid \frac{\partial g}{\partial t} + bF.$$

Thus

$$\frac{\partial g}{\partial s} = \frac{1}{\mid F \mid}\frac{\partial \alpha}{\partial s} - a\frac{F}{\mid F \mid}, \quad \frac{\partial g}{\partial t} = \frac{-1}{\mid F \mid}\frac{\partial \beta}{\partial t} - b\frac{F}{\mid F \mid},$$

and consequently

$$dV = \frac{1}{3}\left[\frac{1}{\mid F \mid}\frac{\partial \alpha}{\partial s}ds,\ -\frac{1}{\mid F \mid}\frac{\partial \beta}{\partial t}dt,\ \frac{F}{\mid F \mid}\right],$$

since the terms $a\frac{F}{|F|}$ and $b\frac{F}{|F|}$ contribute nothing to the triple product.

Integrating along the whole torus $s \in \mathbb{S}^1_1$, $t \in \mathbb{S}^1_2$, we should get the degree $\deg g$ of the Gauss map times the volume $4\pi/3$ of the ball bounded by the unit sphere $\mathbb{S}^2$. Hence, the degree of the Gauss map is provided by the integral along our torus

$$\deg(g : \mathbb{T}^2 \to \mathbb{S}^2) = \frac{1}{(4\pi/3)} \frac{1}{3} \iint dV = \frac{1}{4\pi} \oiint \frac{[\frac{d\alpha}{ds}, \frac{d\beta}{dt}, (\alpha(s) - \beta(t))]}{\mid F(s,t) \mid^3} ds dt,$$

which proves the Gauss integral formula (3) for the linking number of two disjoint closed curves in Euclidean 3-space, $L(\alpha, \beta) = \deg g$. $\square$

The Kronecker characteristic. Kronecker's initial point —whose development led Poincaré to the theory of the degree of the maps, of the vector field indices and to the intersection and liking theories— was the characteristic of three polynomials of two real variables, extending the Sturm characteristic of a pair of polynomials of one real variable. Kronecker's definition was based on the following construction of an integer number associated to $n + 1$ functions of n real variables. We shall consider the case $n = 2$.

Let f, g and h be three real functions of two real variables x, y. Associate to any root (x, y) of the system of equations

$$g(x, y) = 0, \qquad h(x, y) = 0,$$

the sign

$$s(x, y) = \operatorname{sign} \det \begin{vmatrix} \partial g/\partial x & \partial g/\partial y \\ \partial h/\partial x & \partial h/\partial y \end{vmatrix}.$$

To make the root simple, we suppose that the determinant is different from 0.

Take the sum of the signs of all the roots belonging to the domain where $f(x, y) > 0$, and subtract the sum of the signs of the roots belonging to the domain where $f(x, y) < 0$. We suppose there is no root at the boundary curve.

Definition. The *Kronecker characteristic* $K[f, g, h]$ of three functions with no common root for all the three, is the difference between the sum of the signs in the domains $f > 0$ and $f < 0$:

$$K[f, g, h] \; = \sum_{\substack{\{(x,y)\,:\, f\,>\,0,\\ g\,=\,h\,=\,0\}}} s(x, y) \; - \sum_{\substack{\{(x,y)\,:\, f\,<\,0,\\ g\,=\,h\,=\,0\}}} s(x, y)\,.$$

Problem. Prove that *the Kronecker characteristic remains constant when the functions are deformed by a homotopy such that they never have all together a common root. It is also constant if the functions are permuted evenly* (cyclically for 3 functions)*, changing its sign under the odd permutations.*

Solution. Define the map $\varphi : \mathbb{R}^2 \to \mathbb{S}^2$ of the oriented 2-plane $\{(x, y)\}$ to the oriented 2-sphere in oriented Euclidean space $\mathbb{R}^3$ with coordinates (f, g, h),

$$\varphi(x, y) = \frac{F(x, y)}{\mid F(x, y) \mid}, \quad \text{where} \;\; F(x, y) = (f(x, y), g(x, y), h(x, y)) \in \mathbb{R}^3.$$

Choose as ω the North pole, $f = 1$, $g = h = 0$. The Kronecker characteristic is the sum of the signs of the preimages of ω and of $-\omega$: $K[f, g, h] = 2 \deg \varphi$.

10.6 Homotopy groups $\pi_{n+k}(\mathbb{S}^n)$ and cobordisms

We have seen many useful applications of the degree and of the homotopy groups $\pi_n(\mathbb{S}^n)$. The higher dimensional homotopy groups $\pi_{n+k}(\mathbb{S}^n)$ are, in general, unknown, but we shall see now some very important new things in the case $k = 1$.

We already computed the trivial group $\pi_2(\mathbb{S}^1) = 0$. We shall now study the next case, which is much more interesting: The homotopy group $\pi_3(\mathbb{S}^2)$.

We start, as in degree theory, by the study of smooth spheroids, since each element of our group is realisable by a smooth spheroid $\varphi : \mathbb{S}^3 \to \mathbb{S}^2$ (by the Weierstrass polynomial approximation theorem).

Taking a non critical value $\omega \in \mathbb{S}^2$, consider its whole preimage. According to the implicit function theorem, this preimage is a smooth curve α, embedded in $\mathbb{S}^3$.

We shall use the standard Riemannian metric of this 3-sphere, and we fix (arbitrarily) the orientations of both spheres $\mathbb{S}^3$ and $\mathbb{S}^2$.

The 2-planes V of the tangent spaces of $\mathbb{S}^3$, orthogonal to the curve α, are sent isomorphically to the tangent plane $T_\omega\mathbb{S}^2$ by the derivative of φ. Choosing an orienting frame (f_1, f_2) in $T_\omega\mathbb{S}^2$, we lift it to each orthogonal 2-plane V and obtain there an orienting frame (e_1, e_2), sent to (f_1, f_2) by the derivative of φ.

Thus the smooth curve α is framed by these frames (e_1, e_2) of the planes V, orthogonal to the curve. We shall see now that this framed curve contains all homotopic information on the spheroid φ: The spheroid is determined by the framed curve α up to a homotopy.

As it was the case for the 0-dimensional preimages in degree theory, the homotopy relation between the spheroids is equivalent to some interesting equivalence relation between the framed curves, called framed cobordism.

We shall explain it now. But the reasoning is so general that we shall do it for the more general case of the smooth spheroid $\varphi : \mathbb{S}^{n+k} \to \mathbb{S}^n$, with any value k of the dimension excess, making the whole preimage of a non critical value a k-dimensional submanifold of the manifold $\mathbb{S}^{n+k}$, rather than a curve.

This preimage submanifold $\alpha = \varphi^{-1}(\omega)$ is equipped with a smooth normal framing $(e_1(x), \ldots, e_n(x))$ in the n-dimensional space $V(x)$ orthogonal to the

k-dimensional submanifold α at its point x.

Definition. Two orthogonally framed submanifolds A and A' of dimension k of a manifold M^{n+k} are said to be *cobordant framed in the cylinder* $M^{n+k} \times I$ if there exists an orthogonally framed $(k+1)$-dimensional smooth submanifold B of $M^{n+k} \times I$ such that $\partial B = A' - A$ (taking the orientations into account) and at its two boundaries the framings become the sections A and A' of B framed orthogonally in $M^{n+k} \times 0$ and $M^{n+k} \times 1$.

Theorem 9. *Two normally framed smooth submanifolds α and α' of dimension k in $\mathbb{S}^{n+k}$, corresponding to two homotopic spheroids φ and φ', are cobordant framed submanifolds in the cylinder $\mathbb{S}^{n+k} \times I$.*

Every k-dimensional framed submanifold is realisable by a spheroid and every orthogonal framed cobordism provides a homotopy between its boundary spheroids.

The proof of this theorem of Pontryagin is completely parallel to the above study of the case $k = 0$, and we shall not describe all the formal details of this very natural construction.

It reduces the problem of the calculation of the homotopy groups to the questions of the classification of the framed cobordisms, which we have solved in the case $k = 0$ by counting the signs of the framed preimages.

In the case $k = 1$, the preimages are curves, and we have to study their cobordisms.

Theorem 10. *All closed curves are cobordant 1-dimensional manifolds.*

Proof. The pants provide a cobordism between the curves $\mathbb{S}^1 + \mathbb{S}^1$ and $\mathbb{S}^1$, having two connected components and one component, respectively. This implies the cobordism between any finite set of closed curves and one simple circle (which is itself cobordant to zero, being the boundary of a disc). $\square$

This theorem is, however, insufficient for our goals, since it provides no cobordism of framed curves and no cobordism of embedded curves on the corresponding cylinders.

Lemma. *Every knot is cobordant to a standard plane circle, in the sense of the embeddings to the cylinders.*

Proof. To unknot a knot, we can deform it continuously, but we have to pass by simple local crossings a finite number of times. Thus we shall construct a homotopy which joins the curve before the local crossing to the curve after that crossing. Then we have to show that near the moment of crossing, the curve before the crossing is cobordant to the curve after the crossing.

Such a cobordism is shown in Fig. 10.16, and we describe it explicitly:

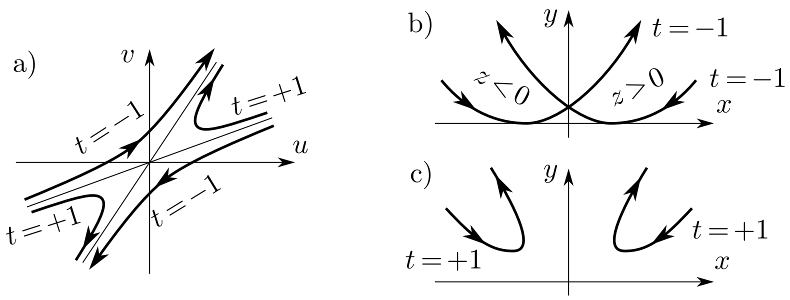


Figure 10.16: A cobordism that eliminates an intersection of the projected space curves.

Consider the following four functions of the two real variables (u, v):

$$x = u, \quad y = v^2, \quad z = v, \quad t = (v - 2u)(u - 2v).$$

The domain of the (u, v)-plane between the two hyperbolas of Fig. 10.16a (defined by the equation $t = (v - 2u)(u - 2v)$ with $-1 \leqslant t \leqslant 1$) is sent by these four functions to a smooth surface in the 4-space belonging to the slice where $-1 \leqslant t \leqslant 1$.

The two branches of the hyperbola $t = -1$ are sent to two curves in the 3-space, with coordinates (x, y, z), whose projections to the (x, y)-plane along the z-direction present an intersection – Fig. 10.16b (such intersections of the projections are the reasons of the knotting of the curves in the 3-space). The same surface in the 4-space represents a cobordism between the 3-space curve of two components $t = -1$ and the 3-space curve of two components $t = +1$, which is the image in $\mathbb{R}^4$ of the two branches of the second hyperbola. The projections to the (x, y)-plane along the z-direction of the two components of the 3-space curve $t = +1$, have no intersection – Fig. 10.16c.

Thus our cobordism unknots the initial curve by elimination of the crossing of the projections of its two parts. In order to kill every crossing of the projected curves, we use this construction a finite number of times. In this way we unknot the curves in $\mathbb{R}^3$ or in $\mathbb{S}^3$.

Once we have unknoted a closed curve (eventually, with several connected components) by a cobordism, we can reduce it to a unique standard plane circle by using the "pants" cobordism which transform $\mathbb{S}^1 \cup \mathbb{S}^1$ into $\mathbb{S}^1$. $\square$

However, to study the homotopy groups, we need the orthogonally framed cobordisms.

Theorem 11. *Every smooth connected cobordism surface with non-void boundary in a 4-manifold can be framed orthogonally.*

Proof. Choose any (generic) orthogonal vector field. It may have some (isolated) zeros. By a homotopy of diffeomorphisms of the surface these zeros can be pushed to the boundary curve (and even to the continuation of the surface outside this curve). As a result we get an orthogonal vector field with no zeros inside the surface.

With this vector field we will construct a field of frames in the planes V orthogonal to the surface, in the following way. Taking the orientation of the surface into account and using the orientation of the ambient 4-space, we orient every Euclidean plane V orthogonal to the surface. Now, we turn each vector of the field by $\pi/2$ in the positive direction, defined by the orientation of each plane V. Thus we get a second orthogonal vector field, and hence we have an orthogonal framing of our surface (whose frames can be chosen to be orthogonal frames of these Euclidean planes). $\square$

To complete the description of the homotopy classes of spheroids, we have to decide which orthogonal framings of the preimage $\varphi^{-1}(\omega)$ of a non critical value $\omega \in \mathbb{S}^2$ determine homotopic spheroid maps $\varphi : \mathbb{S}^3 \to \mathbb{S}^2$.

We shall found an invariant that distinguishes different homotopy classes:

10.6.1 The Hopf Number

Take two non critical values ω and ω' and consider the orthogonally framed preimage curves $\alpha = \varphi^{-1}(\omega)$ and $\alpha' = \varphi^{-1}(\omega')$ in $\mathbb{S}^3$.

Theorem 12. *The linking number $L(\alpha, \alpha')$ does not depend on the choice of the non critical values ω and ω'. It does not depend on the choice of the representative φ of its homotopy class: $L(\alpha, \alpha') = L(\widetilde{\alpha}, \widetilde{\alpha}')$ for any homotopic spheroids φ and $\widetilde{\varphi}$.*

Proof. The independence from the choice of ω and ω' follows from the φ-independence, since one can send any pair of different points ω and ω' of $\mathbb{S}^2$ to any other pair $\widetilde{\omega}$ and $\widetilde{\omega}'$ at the end of a homotopy of diffeomorphisms of the sphere $\mathbb{S}^2$.

If the spheroids φ and $\widetilde{\varphi}$ are homotopic, we can choose a smooth homotopy $\Phi : (I^3 \times I) \to \mathbb{S}^2$ such that $\Phi(t,0) \equiv \varphi(t)$ and $\Phi(t,1) \equiv \widetilde{\varphi}(t)$.

The preimages of the non critical values ω and $\omega' \in \mathbb{S}^2$ under the map Φ are two disjoint oriented bidimensional surfaces $A \subset I^3 \times I$ and $A' \subset I^3 \times I$, which intersect the initial face $I^3 \times \{0\}$ along the orthogonally framed curves γ and γ', and the final face $I^3 \times \{1\}$ along the orthogonally framed curves $\widetilde{\gamma}$ and $\widetilde{\gamma}'$. Theorem 12 follows from the following proposition.

Proposition. *The linking numbers in both faces are equal:*

$$L(\gamma,\gamma') \text{ in } I^3 \times \{0\} \text{ is equal to } L(\widetilde{\gamma},\widetilde{\gamma}') \text{ in } I^3 \times \{1\}$$

Proof. We take two suitable oriented surfaces σ in $I^3 \times \{0\}$ and $\widetilde{\sigma}$ in $I^3 \times \{1\}$, such that $\gamma = \partial\sigma$ and $\widetilde{\gamma} = \partial\widetilde{\sigma}$.

With them, we construct the closed oriented surface $B = \sigma \cup A \cup \widetilde{\sigma}$, in the 4-dimensional product $I^3 \times I$, orienting σ, $\widetilde{\sigma}$ and A so that $\partial\sigma = \gamma$, $\partial\widetilde{\sigma} = -\gamma$

Lemma. *The intersection number of the surface B with the surface A' in the 4-space $I^3 \times I$ equals zero.*

Proof. Indeed, we can slightly deform B to put it in the interior part of the slice $0 \leqslant s \leqslant 1$, and then to translate it along this slice of the 4-space $\mathbb{R}^3 \times I$ far from the surface A'. This homotopy does not change the intersection number of the surfaces B and A', and at the end the deformed surfaces will have no intersection point. Consequently, the intersection number was zero also before the homotopy. The lemma is proved. □

This intersection number can be also computed in other way. The surfaces A and A' have no intersection point, since $\omega \neq \omega'$. So, the intersection number counts the intersection points of the surfaces σ and $\widetilde{\sigma}$ with the surface A'.

Hence, one has to intersect σ and $\widetilde{\sigma}$ with the traces of the surface A' on the faces $s = 0$ and $s = 1$, which are the curves $\gamma' = \varphi^{-1}(\omega')$ and $\widetilde{\gamma}' = \widetilde{\varphi}^{-1}(\omega')$.

The intersection numbers of the surfaces σ and $\widetilde{\sigma}$ with the curves γ' and $\widetilde{\gamma}'$, respectively, in the 3-dimensional faces $s = 0$ and $s = 1$ of the product $I^3 \times I$ are, by definition, the linking numbers

$$(\sigma,\gamma') = L(\gamma,\gamma'), \quad (\widetilde{\sigma},\widetilde{\gamma}') = L(\widetilde{\gamma},\widetilde{\gamma}').$$

In consequence, the vanishing of the intersection number of the surfaces B and A' implies the coincidence of the linking numbers $L(\gamma, \gamma') = L(\widetilde{\gamma}, \widetilde{\gamma}')$ corresponding to the homotopic spheroids φ and $\widetilde{\varphi}$. This proves the above proposition and Theorem 12. □

Thus, we obtain a numerical invariant that distinguishes different homotopy classes: If two spheroids have different linking numbers for the preimages of two non critical values, then these spheroids are not homotopic.

In fact, this invariant is the only one.

Hopf Invariant. The *Hopf number* of the homotopy class of a spheroid is the linking number of the preimage curves (γ and γ') of any two non-critical values of that spheroid,

$$Hopf[\varphi] = L(\gamma, \gamma').$$

10.6.2 Computing the Homotopy Groups $\pi_{n+1}(\mathbb{S}^n)$

Theorem 13. *Two spheroids $\varphi : I^3 \to \mathbb{S}^2$ and $\widetilde{\varphi} : I^3 \to \mathbb{S}^2$ are homotopic if and only if they have equal Hopf numbers. The Hopf invariant can take arbitrary integral values for suitable spheroids:*

$$\pi_3(\mathbb{S}^2) = \mathbb{Z}.$$

Proof. The reasoning is the same as for the case of degree theory. Consider a thin tubular neighbourhood of the preimage curve of a non critical value ω, namely the set $W = \varphi^{-1}(U)$, where U is a small disc around ω in $\mathbb{S}^2$.

Contracting the complement of U in $\mathbb{S}^2$, $\widetilde{U} = \mathbb{S}^2 \setminus U$, to the distinguished point $*$ by a homotopy "along the meridians", we construct a homotopy of our spheroid φ to a "special spheroid" which sends the whole complement of the tubular neighbourhood W of the preimage curve $\gamma = \varphi^{-1}(\omega)$ to the distinguished point $*$.

So, every spheroid is homotopic to some special one (as above), and we only have to find which of the special spheroids are homotopic.

It is sufficient to consider the standard tubular neighbourhood W of the standard plane circle γ in I^3, since we can transform every spheroid to a homotopic spheroid with such a preimage curve (by the orthogonally framed cobordisms theorem).

To each spheroid $I^3 \to \mathbb{S}^2$ (that is, to each map $\mathbb{S}^3 \to \mathbb{S}^2$) there corresponds a framed curve $\varphi^{-1}(\omega)$. Two such spheroids are homotopic if and only if the respective framed curves are framed cobordant (by the Pontryagin theorem, p. 390).

In order to calculate the Hopf invariant in the case of the standard circle, we have to distinguish the different orthogonal framings of the standard circle, which correspond to the different values of the Hopf invariant.

To do it, choose as ω' a point very close to our non critical value ω, in the direction of the tangent vector $f \in T_\omega \mathbb{S}^2$. The preimage vectors e of this vector in the planes V orthogonal to the standard curve γ, determine a framing of γ.

The standard plane circle can be framed by the vectors orthogonal to it, of the plane containing the circle, say, by the outside normals. In this case the linking number $L(\gamma, \gamma')$ vanishes, since one can take the interior plane disc as the surface σ with boundary γ, and it does not intersect the curve γ' at all.

On each plane V orthogonal to γ, we define an angular coordinate ϑ by its vanishing on the vector e (projected by φ_* to f), and using the orientation of the planes V to define the sign of the angles. We obtain the description of every orthogonal framing of the curve γ by the prescription of the direction ϑ of the frame vector in $V(p)$ as a function $\vartheta = \vartheta(p)$ of the point p in γ.

The framing corresponds to the Hopf number k (that is, to the linking number $L(\gamma, \gamma') = k$ of the curves γ and γ') if and only if the increment of the continuous function $\vartheta(p)$ equals $2\pi k$ for one positive turn of the point p along the oriented curve γ.

In this way, we construct spheroids with arbitrary integer values of the Hopf invariant. Since every spheroid is homotopic to one of these standard spheroids (as we have proved above), we get that the Hopf invariant is the only characteristic of the homotopy classes of the spheroids, $\pi_3(\mathbb{S}^2) = \mathbb{Z}$. $\square$

Problem. Prove that $\pi_4(\mathbb{S}^3) = \mathbb{Z}_2$.

Solution. The reasoning is exactly the same as in the case of $\pi_3(\mathbb{S}^2)$: One has to classify the orthogonally framed curves in $\mathbb{S}^4$.

Since in this dimension the knoting of the curves is absent, the problem is simpler than for the Hopf invariant theory: all the problem is to classify the orthogonal framings of the standard plane circle in I^4.

Fixing a standard (trivial) framing, we describe the other framings as continuous paths in the group SO(3), how might be turned the standard frame in the 3-dimensional space orthogonal to γ in I^4.

Knowing that $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$ (Ch. 3, p. 83), we conclude that the homotopy group $\pi_4(\mathbb{S}^3)$ consists of just two elements. These two elements are distinguished by the rotation of the preimages $(f_1(p), f_2(p), f_3(p))$ of the standard frame (e_1, e_2, e_3) of the tangent space $T_\omega\mathbb{S}^3$. Therefore $\pi_4(\mathbb{S}^3) = \mathbb{Z}_2$.

Problem. Calculate the homotopy groups $\pi_{n+1}(\mathbb{S}^n)$, $n \geqslant 3$.

Answer. They are all isomorphic : $\pi_{n+1}(\mathbb{S}^n) = \mathbb{Z}_2$.

Solution. The proof is the same as for $n = 3$, but here one uses the fundamental group stabilisation : $\pi_1(\mathrm{SO}(n)) = \mathbb{Z}_2$, for $n \geqslant 3$.

Problem. Compute the homotopy group $\pi_3(\mathrm{SO}(4))$.

Solution. The first vector fibration over the sphere $\mathbb{S}^3$ with fibre SO(3),

$$\mathrm{SO}(3) \to \mathrm{SO}(4) \to \mathbb{S}^3,$$

provides the homotopy exact sequence

$$\pi_4(\mathbb{S}^3) \to \pi_3(\mathrm{SO}(3)) \to \pi_3(\mathrm{SO}(4)) \to \pi_3(\mathbb{S}^3) \to \pi_2(\mathrm{SO}(3)).$$

We already know four of these homotopy groups (pages 83, 362) :

$$\pi_4(\mathbb{S}^3) = \mathbb{Z}_2, \quad \pi_3(\mathrm{SO}(3)) = \mathbb{Z}, \quad \pi_3(\mathbb{S}^3) = \mathbb{Z}, \quad \pi_2(\mathrm{SO}(3)) = 0\,.$$

From them we get the exact sequence for the unknown group $X = \pi_3(\mathrm{SO}(4))$:

$$\mathbb{Z}_2 \to \mathbb{Z} \to X \to \mathbb{Z} \to 0.$$

Since the image of the leftest homomorphism is 0, we get $\pi_3(\mathrm{SO}(4)) = \mathbb{Z}^2$.

Corollary. *The group* $\pi_3(\mathrm{Spin}(4))$ *is isomorphic to* $\mathbb{Z}^2$*.*

Proof. It follows from the 2-fold covering $\mathrm{Spin}(4) \to \mathrm{SO}(4)$, but we have already seen in quaternion theory (p. 98) that $\mathrm{Spin}(4) = \mathbb{S}^3 \times \mathbb{S}^3$ (providing one more solution of the preceding problem). □

10.6.3 Remarks on Homotopy Groups $\pi_{n+k}(\mathbb{S}^n)$ for $k \geqslant 2$

The cases of higher excess, $\pi_{n+k}(\mathbb{S}^n)$, $k \geqslant 2$, are more difficult, since we need to classify the cobordisms of the k-dimensional manifolds.

Repeating the previous arguments, one can prove the stabilisation (the group does not depend on $n \geqslant n_0(k)$), but even the stabilised groups $\pi_{\infty+k}(\mathbb{S}^\infty)$ are difficult to calculate.

The first attempts by Pontryagin, who invented the cobordism approach, and by his student Rokhlin provided wrong answers for the orthogonally framed cobordisms of the 2-dimensional and 3-dimensional manifolds.

The cobordisms of the unframed theory have been computed; there are even several versions depending on the choices of the oriented or non oriented manifolds and cobordisms.

These cobordism classes form rings, where the addition of manifolds is their disjoint sum and the multiplication is the direct product construction.

The cobordism rings were computed by R. Thom. They are generated by the cobordism classes of the real and complex projective spaces, which verify non trivial relations in these rings. This is one of the most important results of the XXth century mathematics.

Problem. Prove that all closed orientable surfaces belong to the same cobordism class.

Solution. Embed such a surface M in Euclidean space $\mathbb{R}^3$ strictly inside a ball B^3, bounded by a sufficiently big sphere. The embedded surface M bounds some subdomain G^3 of B^3. The complement of G^3 in B^3, $B^3 \setminus G^3$, is a cobordism between the surfaces $\mathbb{S}^2 = \partial B^3$ and $M = \partial G^3$, which proves that they are cobordant (and they are also cobordant to zero, since the sphere $\mathbb{S}^2$ is the boundary of the ball B^3).

An interesting property of the homotopy groups of the spheres is that *all homotopy groups of the spheres are finite, except two families*: $\pi_n(\mathbb{S}^n)$ and $\pi_{4k-1}(\mathbb{S}^{2k})$.

The second family is a generalisation of the Hopf invariant, (that is, the group $\pi_3(\mathbb{S}^2) = \mathbb{Z}$ discussed above) in which the odd-dimensional preimages $\gamma^{2k-1} = \varphi^{-1}(\omega)$, $\gamma'^{2k-1} = \varphi^{-1}(\omega')$ are linked in the sphere $\mathbb{S}^{4k-1}$ of the spheroid $\varphi : I^{4k-1} \to \mathbb{S}^{2k}$ of the group $\pi_{4k-1}(\mathbb{S}^{2k})$.

The linking of two submanifolds X^{ℓ_1} and Y^{ℓ_2} in Z^m occurs when there are generically no intersections ($\ell_1+\ell_2 < m$), but when they become unavoidable in generic one-parameter families of pairs X, Y, that is, when $\ell_1+\ell_2=m-1$ (like for X^1 and Y^1 in Z^3).

In the above example of the spheroids we have $\ell_1=\ell_2=2k-1$ and $m=4k-1$, so that the condition $\ell_1+\ell_2=m-1$ (for the linking of the preimages) holds.

10.7 Helicity of divergence-free vector fields

The linking numbers of the curves in 3-space have very important physical applications in magneto-hydrodynamics, where the Hopf invariant generalised to vector fields is known under the name of *helicity*.

To simplify some formulas, we explain here the main idea for the case of the oriented 3-sphere. There exist similar theories for other cases; for instance, for the oriented Euclidean 3-space.

Let v be a smooth vector field on $\mathbb{S}^3$. The standard oriented volume element 3-form τ on $\mathbb{S}^3$ and the field v determine the differential 2-form ω_v whose value on two vectors ξ, η, tangent to the sphere at the same point x, is equal to the oriented volume of the parallelepiped generated by the vectors ξ, η and $v(x)$ in $T_x\mathbb{S}^3$: $\omega_v(\xi,\eta)=\tau(\xi,\eta,v(x))$.

The field v is said to be *divergence free* if this form ω_v is closed: $d\omega_v=0$.

It is equivalent to the condition that the flow of the vector field v preserves the volumes of the domains: $\iiint_{g^tD}\tau=\iiint_D\tau$ for any domain $D\subset\mathbb{S}^3$ and any time moment t.

It is also equivalent to the existence of a *vector potential* for v, which is a 1-form α on $\mathbb{S}^3$ whose exterior derivative is ω_v:

$$\omega_v=d\alpha.$$

This condition means that the integral of the 1-form α along the boundary curve of any oriented surface σ is equal to the integral of the form ω_v along this surface,

$$\iint\limits_{\sigma}\omega_v=\int_{\partial\sigma}\alpha.$$

In other words, the circulation of the vector potential along the boundary curve is equal to the flux of the mass τ through the surface σ.

We suppose our vector field v is divergence free.

Definition. The *helicity* of a divergence-free vector field v is the number

$$H[v] = \iiint\limits_{\mathbb{S}^3} \omega_v \wedge \alpha,$$

where α is a vector potential for v: $\omega_v = d\alpha$.

This triple integral does not depend on the particular choice of the potential α. Indeed, for another potential $\widetilde{\alpha}$ we would have $\widetilde{\alpha} - \alpha = df$ for some function f. Therefore,

$$\omega_v \wedge \widetilde{\alpha} - \omega_v \wedge \alpha = \omega_v \wedge (df) = d(\omega_v f),$$

and by the Stokes formula

$$\iiint\limits_{\mathbb{S}^3} (\omega_v \wedge \widetilde{\alpha} - \omega_v \wedge \alpha) = \iint\limits_{\partial\mathbb{S}^3} (\omega_v f) = 0,$$

which proves the irrelevance of the choice of α.

The main point is the geometric interpretation of the helicity integral H in terms of the linking of the curves tangent to the vector field v.

Let x be a point of $\mathbb{S}^3$. The curve φ of the field v, starting at x, is defined as the solution $\varphi : [0, s] \to \mathbb{S}^3$ of the differential equation

$$\frac{d\varphi}{dt} = v(\varphi(t))$$

with initial condition $\varphi(0) = x$.

It is a smooth parametrised curve depending on the "time" s, but it is not closed. So, to close it, we add a short path returning from $\varphi(s)$ to x. We generate in this way a closed curve $\widehat{\varphi}_s$ in $\mathbb{S}^3$, which depends on s (see Fig. 10.17).

Similarly, we close the curve $\psi : [0, t] \to \mathbb{S}^3$ of the field v, starting at a point $\psi(0) = y \in \mathbb{S}^3$, to get a closed curve $\widehat{\psi}_t$ in $\mathbb{S}^3$.

If these two curves have no common point, we consider their linking number

$$L(\widehat{\varphi}_s, \widehat{\psi}_t) = L(x, s; y, t).$$

If one covers a closed curve twice, its linking number would double its value. So, one expects that the function $L(x, s; y, t)$ behaves (for large lengths s and t of the curves) approximately as a linear function of s and of t.

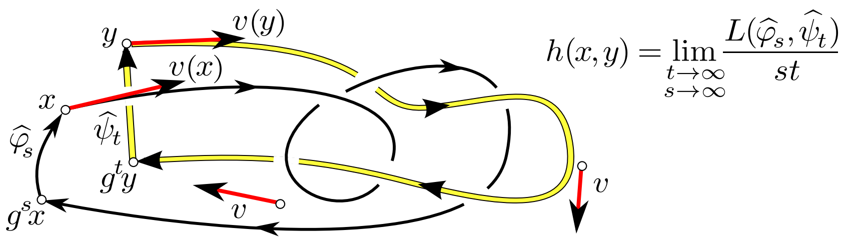


Figure 10.17: The local helicity of a vector-field by its curves $\widehat{\varphi}$, $\widehat{\psi}$, for the starting points x and y.

Hence, one defines the *local helicity of the vector field* v at the points x and y, as the limit

$$h(x,y) = \lim_{\substack{s \to \infty \\ t \to \infty}} \frac{L(x,s;y,t)}{st}. \tag{4}$$

The existence of this limit is not evident, but it is proved to exist for almost any pair of initial points x and y and for generic short return paths, used to close the curves. The limit is also independent of the choices of these short paths, whose change produces an increment in the linking number, which is asymptotically much smaller than the large product st.

In the case of a vector field directed along the preimage curves of the points of $\mathbb{S}^2$, in the sphere $\mathbb{S}^3$, for a spheroid $\varphi : \mathbb{S}^3 \to \mathbb{S}^2$, the local helicity is generically independent of the points x and y (whose images $\varphi(x)$ and $\varphi(y)$ should only be non critical values).

However, the ergodic theory limit (4) is well defined for any divergence free vector field, and if this vector field is not provided by a map to $\mathbb{S}^2$, the values of the local helicity functions may be different for different choices of the initial points x and y.

We can eliminate the second point y by making it infinitely close to x, that is, choosing instead of the point y a tangent vector f to $\mathbb{S}^3$ at x, which should not be proportional to $v(x)$.

In this case we get a new "local helicity function" on $\mathbb{S}^3$, whose value $\widehat{h}(x)$ measures the asymptotic speed of rotation of the preimages e of the vector f in the planes V orthogonal to the curve φ at its different points.

Thus, we have defined two local helicity functions,

$$h(\cdot,\cdot) : \mathbb{S}^3 \times \mathbb{S}^3 \to \mathbb{R}, \quad \widehat{h}(\cdot) : \mathbb{S}^3 \to \mathbb{R}.$$

The topological description of the helicity integral H is the following

Theorem 14. *The mean values of the local helicity functions are equal to the helicity integral,*

$$H[v] = \frac{\iiiint\!\!\iint_{\mathbb{S}^3\times\mathbb{S}^3} h(x,y)d\tau(x)d\tau(y)}{\iiiint\!\!\iint_{\mathbb{S}^3\times\mathbb{S}^3} 1 d\tau(x)d\tau(y)}.$$

Many details on helicity theory and on its applications to the topological studies of the magnetic fields are described in the elementary textbook [41].

The following important question remains open: *Is the helicity of a vector field (and are the local helicity functions) a topological invariant of the field (with respect to the homeomorphisms* $\mathbb{S}^3 \to \mathbb{S}^3$ *preserving the* τ*-volumes)*?

They are invariant with respect to such volume-preserving diffeomorphisms (by the very definitions), but the smooth structure was used to define them, and the homeomorphisms are not acting on the vector fields. So, one has to explain the meaning of the problem.

Definition. Two smooth vector fields v and $\widetilde{v}$, on two smooth manifolds M and $\widetilde{M}$, are called *topologically equivalent* if there exists a homeomorphism $f : M \to \widetilde{M}$ which sends the flow $\{g^t\}$ of v to the flow $\{\widetilde{g}^t\}$ of $\widetilde{v}$:

$$\widetilde{g}^t(f(x)) = f(g^t x) \quad \forall x \in M, \forall t.$$

The topological invariance statement means the equality of the corresponding object for topologically equivalent vector fields.

No one knows any example of different helicities $H[\widetilde{v}] \neq H[v]$ for topologically equivalent smooth fields v and $\widetilde{v}$, but the equality of their helicities is not proved.

Chapter 11

Homology

Homology theory is the most elaborated part of topology. In many cases, it provides the easiest calculations and the most important numerical invariants of manifolds, vector fields and many other geometrical objects (similar to the genus of an orientable surface, to the Euler characteristic, to the degree of a map and to the linking numbers). In the approach to homology theory presented in this chapter, we try to explain the main ideas of the theory and to show "how it works", focusing on explicit computations and applications, using the basic properties of homologies, and avoiding technical details in the proofs and in the formal construction of the theory.

We introduce and compute the homology of a smooth manifold M. But it is worth to have in mind that the same definition can also be applied to a manifold with boundary, a singular algebraic variety, or even to an arbitrary "topological space".

11.1 Chain complexes and homology groups

In Ch. 2 (p. 46), we have seen that the number of vertices v, of edges e and of faces f of a polyhedron determine its Euler characteristic, $\chi = v - e + f$, which depends only on the topological nature of the surface. It is $\chi = 2$ if the polyhedron is homeomorphic to the sphere and $\chi = 0$ for a polyhedron homeomorphic to the torus.

We shall use here the vertices, edges and faces of regular polyhedra in a more elaborated way. Let G_f, G_e and G_v be the respective free abelian groups of oriented faces, edges and vertices of a tetrahedron (or of a dodecahedron), considered in Sect. 6.5 (pp. 195-196). The boundary homomorphisms of these groups $\partial_f : G_f \to G_e$, $\partial_e : G_e \to G_v$ and $\partial_v : G_v \to 0$, form the sequence

$$0 \xrightarrow{i} G_f \xrightarrow{\partial_f} G_e \xrightarrow{\partial_e} G_v \xrightarrow{\partial_v} 0$$

(with i the inclusion of the zero 2-chain), which satisfies the property

$$\partial_f \circ i = 0\,, \quad \partial_e \circ \partial_f = 0\,, \quad \partial_e \circ \partial_v = 0\,.$$

(indeed, in Ch. 6 we have proved that the boundary of the boundary of a convex polyhedron of any dimension is zero (empty): $\partial_{m-1} \circ \partial_m = 0$).

These equalities mean the following inclusions of abelian subgroups

$$\{0\} \subset \operatorname{Ker} \partial_f\,, \quad \operatorname{Im} \partial_f \subset \operatorname{Ker} \partial_e\,, \quad \operatorname{Im} \partial_e \subset \operatorname{Ker} \partial_v\,.$$

The quotient groups

$$\operatorname{Ker} \partial_f / \{0\}\,, \quad \operatorname{Ker} \partial_e / \operatorname{Im} \partial_f\,, \quad \operatorname{Ker} \partial_v / \operatorname{Im} \partial_e$$

of the subgroups of chains without boundary ($\operatorname{Ker} \partial_*$) modulo the subgroups of boundaries ($\operatorname{Im} \partial_*$) are called *homology groups* of the above sequence of homomorphisms. We note them by H_f, H_e and H_v.

Observe that for the vertices we always have $\operatorname{Ker} \partial_v = G_v$.

A closed oriented path along the edges of a polyhedron can be cyclically gone through because it has not end-points (no boundary). For this reason the closed oriented paths and the formal linear combinations of them are called *cycles*. Let's call *trivial cycle* the boundary of any face of a polyhedron. Observe that the boundary of any sum of faces with multiplicities is a cycle (see Fig. 11.1 where the double arrow on one edge of the chain $\partial(A - B)$ means that this edge has multiplicity 2).

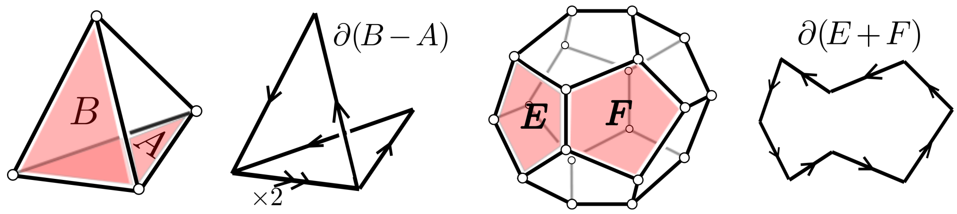


Figure 11.1: Cycles of two regular polyhedra.

Tetrahedron homology groups. Let us calculate the homology groups of the tetrahedron T (Fig. 11.1). Its sequence of boundary homomorphisms is

$$0 \xrightarrow{i} \mathbb{Z}^4 \xrightarrow{\partial_f} \mathbb{Z}^6 \xrightarrow{\partial_e} \mathbb{Z}^4 \xrightarrow{\partial_v} 0.$$

First, although the boundary of each face is not zero, the sum of all faces with multiplicity 1 (giving the oriented tetrahedron) has boundary zero. Thus any multiple of this sum has boundary zero. Hence $\operatorname{Ker}\partial_f \approx \mathbb{Z}$ and then $\operatorname{Im}\partial_f \approx \mathbb{Z}^3$. Next, any closed path through the edges of T is a combination of the four trivial cycles. However, they are not independent because their sum is the boundary of T, which is zero (each edge appears twice with opposite orientations). So $\operatorname{Ker}\partial_e \approx \mathbb{Z}^3$ and then $\operatorname{Im}\partial_e \approx \mathbb{Z}^3$. Finally, these results together with $\operatorname{Ker}\partial_v = \mathbb{Z}^4$ provide the homology groups of T:

$$H_f(T) = \mathbb{Z}/\{0\} = \mathbb{Z}\,, \quad H_e(T) = \mathbb{Z}^3/\mathbb{Z}^3 = \{0\}\,, \quad H_v(T) = \mathbb{Z}^4/\mathbb{Z}^3 = \mathbb{Z}\,.$$

Dodecahedron homology groups. The dodecahedron D (Fig. 11.1) has the following sequence of boundary homomorphisms:

$$0 \xrightarrow{i} \mathbb{Z}^{12} \xrightarrow{\partial_f} \mathbb{Z}^{30} \xrightarrow{\partial_e} \mathbb{Z}^{20} \xrightarrow{\partial_v} 0.$$

Using exactly the same reasonings as for the tetrahedron we get that $\operatorname{Ker}\partial_f \approx \mathbb{Z}$ (implying $\operatorname{Im}\partial_f \approx \mathbb{Z}^{11}$), $\operatorname{Ker}\partial_e \approx \mathbb{Z}^{11}$ (implying $\operatorname{Im}\partial_e \approx \mathbb{Z}^{19}$) and $\operatorname{Ker}\partial_v = \mathbb{Z}^{20}$. Hence the homology groups of the dodecahedron are

$$H_f(D) = \mathbb{Z}/\{0\} = \mathbb{Z}\,, \quad H_e(D) = \mathbb{Z}^{11}/\mathbb{Z}^{11} = \{0\}\,, \quad H_v(D) = \mathbb{Z}^{20}/\mathbb{Z}^{19} = \mathbb{Z}\,,$$

which are isomorphic to those of the tetrahedron.

Exercise. Verify that the homology groups of the other regular polyhedra (and of any convex polyhedron) are isomorphic to those of the tetrahedron.

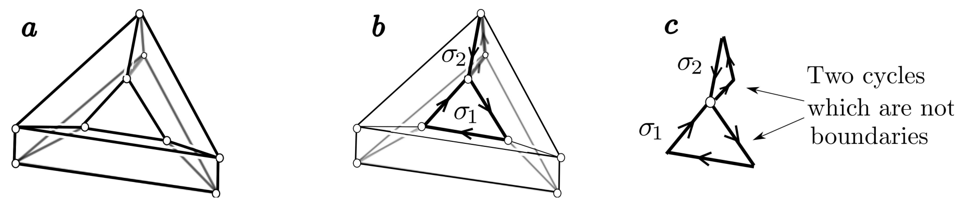


Figure 11.2: Two non boundary cycles on a polyhedral torus.

Homology groups of toric polyhedra. The groups of faces, edges and vertices of the polyhedron P of Fig. 11.2 *a* are $G_f \approx \mathbb{Z}^9$, $G_e \approx \mathbb{Z}^{18}$, and $G_v \approx \mathbb{Z}^9$. Hence its sequence of boundary homomorphisms is

$$0 \xrightarrow{i} \mathbb{Z}^{9} \xrightarrow{\partial_f} \mathbb{Z}^{18} \xrightarrow{\partial_e} \mathbb{Z}^{9} \xrightarrow{\partial_v} 0.$$

The same arguments as above lead to $\operatorname{Ker}\partial_f \approx \mathbb{Z}$ and $\operatorname{Im}\partial_f \approx \mathbb{Z}^8$. As above, the group generated by the 9 trivial cycles has rank 8 in $\mathbb{Z}^{18}$. However, the oriented "parallel" and "meridian" depicted in Fig. 11.2 *b* and 11.2 *c*, are two cycles that are not linear combinations of trivial cycles (that is, they are not boundaries of any linear combination of the faces). Thus the rank of $\operatorname{Ker}\partial_e$ is equal to $8 + 2 = 10$, and hence $\operatorname{Ker}\partial_e \approx \mathbb{Z}^{10}$ (the reader can take other oriented parallel (or meridian) and verify that it is obtained from the parallel $\pm\sigma_1$ (or meridian $\pm\sigma_2$) by adding some trivial cycles). We obtain thus that $\operatorname{Im}\partial_e \approx \mathbb{Z}^8$. These results together with $\operatorname{Ker}\partial_v \approx \mathbb{Z}^9$ provide

$$H_f(P) = \mathbb{Z}\,, \qquad H_e(P) = \mathbb{Z}^{10}/\mathbb{Z}^8 = \mathbb{Z}^2\,, \qquad H_v(P) = \mathbb{Z}^9/\mathbb{Z}^8 = \mathbb{Z}\,.$$

The reader can verify that the homology groups of the toric polyhedra depicted in Fig 2.8 (p. 46) are isomorphic to those of P.

As for the Euler characteristic, the homology groups of a polyhedron depend only on the topological nature of the surface (not on its polyhedral realisation). So, the isomorphic homology groups of the regular polyhedra are also isomorphic to the homology groups of any polyhedron homeomorphic to the sphere. The same is true for the isomorphism of the homology groups of the polyhedra homeomorphic to the torus.

We explain now how to associate the above homology groups to the sphere and to the torus respectively.

A homeomorphism of a convex polyhedron P (say, the tetrahedron or the dodecahedron) to the sphere induces a partition of the sphere into oriented curvilinear faces, curvilinear edges and vertices, respecting the orientation of the boundaries (Fig. 11.3).

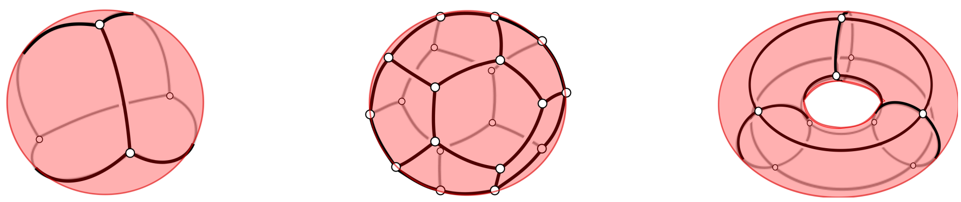


Figure 11.3: Partition of surfaces into curvilinear polygons.

The groups of faces, of edges and of vertices of the sphere (viewed as this curvilinear polyhedron) are isomorphic to those of the polyhedron P, and its boundary homomorphisms are induced from those of P. Hence we have isomorphic homology groups: $H_f(\mathbb{S}^2) = \mathbb{Z}$, $H_e(\mathbb{S}^2) = 0$, $H_v(\mathbb{S}^2) = \mathbb{Z}$. Similarly, a homeomorphism of the polyhedron of Fig. 11.2 *a* to the torus induces a curvilinear polygonal partition of the torus (Fig. 11.3) providing its homology groups $H_f(\mathbb{T}^2) = \mathbb{Z}$, $H_e(\mathbb{T}^2) = \mathbb{Z}^2$, $H_v(\mathbb{T}^2) = \mathbb{Z}$.

Homology groups of any manifold. To generalise the construction of these examples to any n-dimensional manifold M, in order to define and compute its homology groups, one partitions M into oriented "bricks" of different dimensions: n-dimensional curvilinear polyhedra, with all their curvilinear k-faces, $0 \leqslant k \leqslant n-1$. As above, the group of k-faces of M consists of all sums of curvilinear k-faces with integral multiplicities. It is called the *group of k-dimensional chains* (of that partition) of M and is denoted by $C_k(M)$.

Here, a k-dimensional curvilinear polyhedron is a homeomorphism $f : P \to M$ of an oriented rectilinear convex polyhedron $P \subset \mathbb{R}^k$ (see Section 6.5). So, the group of k-dimensional chains $C_k(M)$ of a partition of M (whose generators are the k-dimensional faces of that partition) is a finitely generated subgroup of the infinitely non-countable generated group $\{c^k\}_M$ of singular chains, formed by all finite linear combinations of any possible curvilinear polyhedra*.

To be rigorous, we shall identify bricks which differ only by the choice of parametrisation f (roughly speaking, we identify them with their common geometric image). Moreover, we assume that P is always one and the same rectilinear polyhedron (for example, simplex or cube). We should (but will not) precise the rules to glue the bricks to cover the manifold M.

Thus, we have a sequence of boundary homomorphisms

$$\dots C_k(M) \xrightarrow{\partial_k} C_{k-1}(M) \xrightarrow{\partial_{k-1}} \dots \to C_1(M) \xrightarrow{\partial_1} C_0(M) \xrightarrow{\partial_0} 0$$

that has the property $\partial^2 = 0$, namely, $\partial_{k-1}\partial_k = 0$ for any k (pp. 211-212).

A sequence of homomorphisms with this important semi-exactness property "$\partial^2 = 0$" is called a *chain complex*.

So the sequence of boundary homomorphisms is a chain complex.

Definition. A k-chain $c \in C_k(M)$ is called *cycle in M* if its boundary vanishes:

$$\partial_k c = 0, \qquad c \in \operatorname{Ker} \partial_k.$$

*Note that in the subgroup $C_k(M)$ of $\{c^k\}_M$ the parametrisations f must be homeomorphisms, while in $\{c^k\}_M$ it is possible to have continuous maps f whose image is non homeomorphic to P, it can even have dimension less than k.

The subgroup $\operatorname{Ker}\partial_k$ is called *group of k-cycles of M*. The usual notation is

$$Z_k(M) := \operatorname{Ker}\partial_k \ (\subseteq C_k(M)).$$

Definition. A k-chain $b \in C_k(M)$ is called a *boundary in M* if it is the boundary of some $(k+1)$-chain: $b \in \operatorname{Im}\partial_{k+1} \subseteq C_k(M)$. The boundaries form a subgroup of $C_k(M)$, called the *group of k-boundaries in M*.

The usual notation is

$$B_k(M) := \operatorname{Im}\partial_{k+1} \ (\subseteq C_k(M)).$$

The $\partial^2 = 0$ property implies the inclusion of commutative groups

$$B_k(M) \subseteq Z_k(M),$$

since every boundary is a cycle.

Definition. The *kth homology group* of M is the quotient group

$$H_k(M) = Z_k(M)/B_k(M)$$

of the k-cycles modulo the k-boundaries.

It is very important that the groups of polyhedral chains $C_k(M)$ are finitely generated because this allows to make computations.

Remark from algebra The success of homology theory in several areas of mathematics and physics led the mathematicians to make the algebraic study of chain complexes of arbitrary abelian groups that not necessarily come from geometry or topology. Thus the sequence of boundary homomorphisms of our groups of curvilinear k-faces enters in the following general algebraic setting.

Definition. A sequence C of homomorphisms of commutative groups

$$\cdots \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} \cdots C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

that satisfies $\partial_{k-1} \circ \partial_k = 0$ for any k is called *chain complex*. The homomorphisms ∂_k are called *boundary homomorphisms* or *boundary operators*.

The quotient group $H_k(C) := \operatorname{Ker}\partial_k/\operatorname{Im}\partial_{k+1}$ is called the k-th *homology group* of the chain complex C. The elements of $\operatorname{Ker}\partial_k$ are called *k-cycles* and those of $\operatorname{Im}\partial_{k+1}$ *k-boundaries*.

To save paper, we denote the boundary operators ∂_k simply by ∂.

In this general definition it is also possible to replace the groups by rings, vector spaces or other algebraic structures. The homomorphisms of the sequence must preserve the corresponding algebraic structures. Note that algebraists kept topological terminology "boundary homomorphisms", "cycles" and "boundaries".

The group of coefficients. From this general point of view a natural generalisation of our groups of chains of dimension k in M, say $C_k(M)$ (or $\{c^k\}_M$), consist of considering all sums of curvilinear k-faces with coefficients belonging to an arbitrary commutative group G. In this case the homology groups are noted $H_k(M, G)$. The usual choice of the group of coefficients will be the additive group of integers $\mathbb{Z}$ —in this case the homologies are called "integral homologies"— it would be more pedantic to write $C_k(M, \mathbb{Z})$, but we will write simply $C_k(M)$ in this case.

The other frequent choices are the coefficients 0 and 1 forming $\mathbb{Z}_2$ (leading to the $\mathbb{Z}_2$-homology), or one uses the coefficients of $\mathbb{Z}_p$, or of the additive group of the rational numbers $\mathbb{Q}$, or of the real numbers $\mathbb{R}$, and some times of the complex numbers $\mathbb{C}$ (leading to the homologies with the corresponding prefix: $\mathbb{R}$-homologies and $\mathbb{R}$-chains and so on).

In all these cases there is a well defined boundary of a k-chain, which is a $(k-1)$-chain, for the same group of coefficients.

Different groups of chains. We see that the fundamental necessary notion to construct homology groups is that of chain complex: One partitions the manifold into bricks to get the chain groups and the chain complex (of that partition), which bring the homology groups (with the chosen coefficients). Some difficulty of this general construction is the possibility to start with different groups of chains (coming from different partitions). To make it easy, one usually fixes one type of oriented brick. For example, a usual choice is to take, for any $k \in \mathbb{N}$, only k-dimensional simplices (or cubes) P in $\mathbb{R}^k$. Let us consider the simplicial chains.

11.1.1 Triangulations and simplicial homology

A *triangulation* (or *simplicial partition*) of a manifold M is a partition of it into simplices satisfying the following two conditions:

(i) M is decomposed into a finite number of oriented simplices of different dimensions, including with every simplex its faces of all dimensions.

(ii) Either the intersection of any two simplices is empty, or one of them is a face of the other, or they have one common face.

These conditions guarantee that the manifold is homeomorphic to a rectilinear polyhedron in some Euclidean space.

The formal linear combinations of the k-dimensional simplices of a triangulation are called *simplicial k-chains of M*.

Observe that the orientation of a k-simplex $\Delta^k = (a_0, \dots, a_k)$ depends only on the ordering of its vertices. Namely, the ordering $a_0, a_1, \dots, a_k$ orients positively the simplex Δ^k if the frame formed by the vectors $\mathbf{e}_1 = \overrightarrow{a_0 a_1}, \dots,$ $\mathbf{e}_k = \overrightarrow{a_0 a_k}$ orients Δ^k correctly. The $(k-1)$-face of Δ^k opposite to the vertex a_j is the $(k-1)$-simplex Δ_j^{k-1} obtained from Δ^k by eliminating the vertex a_j, $\Delta_j^{k-1} = (a_0, \dots, a_{j-1}, a_{j+1}, \dots, a_k)$.

It is easy to verify that the oriented boundary of Δ^k is given by

$$\partial \Delta^k = \sum_{j=0}^{k} (-1)^j \Delta_j^{k-1} \,. \tag{1}$$

The cases of the 1-simplex and the 2-simplex are shown in Fig. 11.4. The reader can verify the case of the tetrahedron.

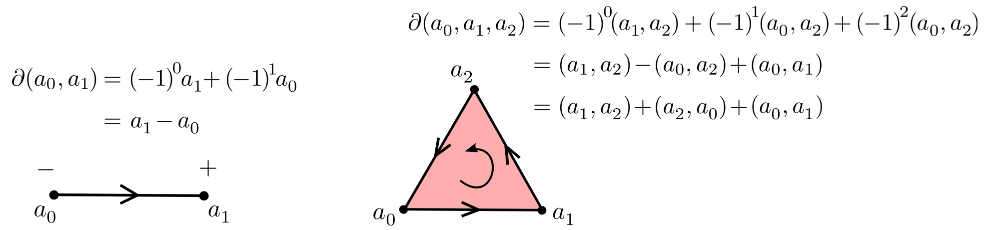


Figure 11.4: Boundary orientation of simplices.

An advantage of using simplicial chains is that once an ordering of its vertices is chosen the orientations and oriented boundaries can be defined and computed in a canonical purely combinatorial way.

For the torus $M = \mathbb{S}^1 \times \mathbb{S}^1 = \mathbb{T}^2$ one can easily find a triangulation with 18 triangles, for example, subdividing each face of the polyhedron of Fig. 11.2 into two triangles. The chain complex of such triangulation is

$$0 \xrightarrow{i} \mathbb{Z}^{18} \xrightarrow{\partial_f} \mathbb{Z}^{27} \xrightarrow{\partial_e} \mathbb{Z}^9 \xrightarrow{\partial_v} 0,$$

whose homology groups are isomorphic to those obtained above for the polyhedron of Fig. 11.2. It is not too difficult to prove that other triangulations of the torus provide isomorphic homology groups.

Similarly, *all different triangulations of a manifold M lead to isomorphic homology groups*, called the *simplicial homology groups* of M.

Generalised triangulations. One can prove that the minimal number of triangles for a triangulation of a torus is 14. However, a smaller number of "triangles" can be obtained if one considers "generalised triangulations" allowing some of the faces of the involved simplices to be identified (Fig. 11.5). Namely, condition (ii) for triangulations (used to prove some theorems and to ensure that the partition "looks like a polyhedron") can be banished to compute homology groups – it does not affect the result.

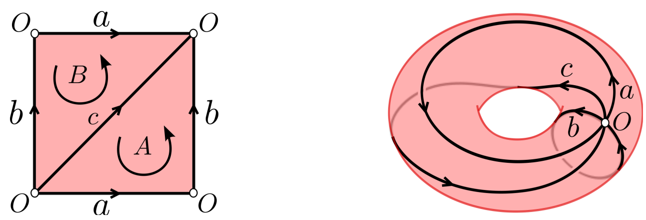


Figure 11.5: A triangulation of a two-dimensional torus, having 6 simplices.

Example. The triangulation of the torus described in Fig. 11.5 is not induced from a homeomorphism of a polyhedron: the two triangles intersect along their three edges and the three vertices of both triangles are identified to the point O. It contains one 0-dimensional simplex O, three 1-dimensional simplices a, b and c, and two 2-dimensional simplices A and B.

For this "triangulation" the spaces of simplicial chains have the following dimensions (say, for $\mathbb{R}$-coefficients) :

$$\dim C_0 = 1, \quad \dim C_1 = 3, \quad \dim C_2 = 2.$$

The boundary operators, called "differentials", act in the following way:

$$\partial A = a + b - c, \quad \partial B = c - a - b,$$

$$\partial a = 0, \quad \partial b = 0, \quad \partial c = 0, \quad \partial O = 0.$$

So, the 2-chain $\alpha A + \beta B$ is a 2-cycle if ($\alpha - \beta = 0$ and $\beta - \alpha = 0$), but it is never a boundary (unless it is itself zero). Hence, the 2-homology group

$$H_2(\mathbb{T}^2, \mathbb{Z}) \approx \mathbb{Z}$$

is generated by the "fundamental cycle" $A + B$ of the torus.

To find the 1-dimensional homology group, we consider all the 1-chains, $pa + qb + rc$, with integral coefficients p, q, r. Their boundaries are all zero, and hence all the 1-chains are cycles:

$$Z_1(\mathbb{T}^2, \mathbb{Z}) = C_1(\mathbb{T}^2, \mathbb{Z}) \approx \mathbb{Z}^3.$$

On the other side, the subgroup B_1 of all 1-boundaries

$$\partial(uA + vB) = (u - v)a + (u - v)b - (u - v)c\,,$$

clearly generated by the element $a + b - c = \partial A$ (with coefficients $p = q = 1$, $r = -1$), has rank one in the group of 1-cycles Z_1.

Therefore, the 1-homology group is isomorphic to $\mathbb{Z}^2$:

$$H_1(\mathbb{T}^2, \mathbb{Z}) = Z_1/B_1 \approx \mathbb{Z}^3/\mathbb{Z} = \mathbb{Z}^2.$$

The classes of the cycles a and b can be chosen as the pairs of generators of the homology group $H_1(\mathbb{T}^2, \mathbb{Z})$.

With the same reasoning we prove that for our triangulation

$$H_1(\mathbb{T}^2, \mathbb{R}) \approx \mathbb{R}^2, \quad H_1(\mathbb{T}^2, \mathbb{Z}_p) \approx (\mathbb{Z}_p)^2, \quad H_1(\mathbb{T}^2, \mathbb{Q}) \approx \mathbb{Q}^2, \quad H_1(\mathbb{T}^2, \mathbb{C}) \approx \mathbb{C}^2.$$

11.1.2 Homologous cycles and homology classes

The coset of a k-dimensional cycle $\sigma \in Z_k(M)$ with respect to the subgroup of boundaries $B_k(M)$ is the equivalence class, denoted by $[\sigma]$, formed by all cycles of the form $\sigma + \partial D$, where D is any $(k+1)$-dimensional chain. So the quotient group $H_k(M)$ is formed by classes $[\sigma]$, called *homology classes*.

We shall say that a cycle σ is *homologous to zero* if it belongs to the zero class, formed by all k-boundaries ($\sigma = \partial D$ for some $(k+1)$-chain D), and that two k-dimensional cycles α and β are *homologous* if their difference is a boundary ($\beta = \alpha + \partial D$ for some $(k+1)$-dimensional chain D), that is, if they belong to the same homology class, $[\alpha] = [\beta]$. Since this "homology relation" between α and β is provided by the existence of such a $(k+1)$-dimensional chain D, one often says that "D is a homology between α and β".

Any oriented k-dimensional closed submanifold N can be considered as a k-cycle: it is the formal sum of the k-simplices forming N (taken with multiplicity 1) for a chosen triangulation. The homology class $[N]$ represented by this cycle is called the *fundamental class* of the submanifold N.

Since the generators of the group $H_k(M)$ are classes whose cycles are not boundaries, in order to compute $H_k(M)$ it is often convenient to recognise independent concrete cycles. For example, for the toric polyhedron P of Fig. 11.2 the classes $[\sigma_1]$ and $[\sigma_2]$ of the nontrivial cycles σ_1 and σ_2 generate $H_e(P) = \mathbb{Z}^2$. In the above example, the cycles a and b generate the group $H_1(\mathbb{T}^2) = \mathbb{Z}^2$ (the cycles $a+b$ and $a-b$ are also generators), the cycle $a+b-c = \partial A$ generates the boundary group Z_1, and the group $H_2(\mathbb{T}^2) = \mathbb{Z}$ is generated by the fundamental class $[A+B]$.

Remark. It is natural to consider "homologous" chains that are not cycles (Fig. 11.6).

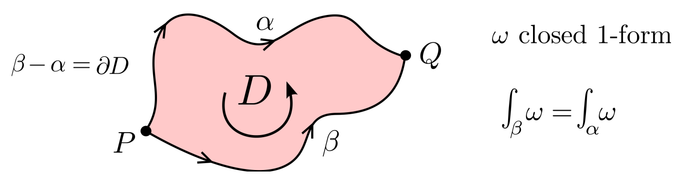


Figure 11.6: The integrals of a closed form over homologous chains.

For example, by Stokes formula, the integrals of a closed k-form ω over two k-dimensional chains α and β are equal provided that the difference of these chains is the boundary of some $(k+1)$-chain D. That is, if $d\omega = 0$ and $\beta - \alpha = \partial D$, then

$$\int_\alpha \omega = \int_\beta \omega \,.$$

Namely,

$$\int_\beta \omega - \int_\alpha \omega = \int_{\beta-\alpha} \omega = \int_{\partial D} \omega \overset{Stokes}{=} \int_D d\omega = 0 \quad \Longrightarrow \quad \int_\beta \omega = \int_\alpha \omega.$$

Poincaré called *homologous* two such chains α and β. This equality of integrals, very useful in physics, was one of the first examples in which the "homology relation" between chains showed to be important, and which, among other things, lead Poincaré to create Homology theory.

We shall prove the following theorem by recognising homologous cycles.

Theorem. *The zeroth homology group of a manifold M with b_0 connected components is $H_0(M) = \mathbb{Z}^{b_0}$.*

Proof. It suffice to prove the statement, $H_0 \approx \mathbb{Z}$, for each connected component because each connected component can be treated separately. The group C_0 of zero-chains is formed by all combinations of the N points of the triangulation, $n_1\alpha_1 + \cdots + n_N\alpha_N$. Fix a point $\alpha = \alpha_j$. Each vertex α_k is homologous to α because it can be connected to α by a path σ formed by oriented 1-dimensional simplices, that is, $\alpha_k - \alpha = \partial\sigma$. Hence all vertices belong to the homology class of α, $[\alpha_k] = [\alpha]$. It generates the group $H_0(M)$ because $[n_1\alpha_1 + \cdots + n_N\alpha_N] = n_1[\alpha_1] + \cdots + n_N[\alpha_N] = (n_1 + \cdots + n_N)[\alpha]$, proving that $H_0(M) = \mathbb{Z}$. $\square$

11.2 Five essential properties of homologies

For a compact manifold M (possibly with boundary) each simplicial homology group $H_k(M, \mathbb{Z})$ is commutative with a finite number of generators.

In the case of the torus $\mathbb{T}^2$, it is not too difficult to prove that other natural choices of the chains lead to the same homology groups. For instance, one can start from the cubic chains (leading to *cubic* homology), or one can start from the infinite dimensional spaces $\{c^k\}_M$ of singular chains along which we have integrated the k-forms (leading to *singular* homology) or from the cellular chains (leading to *cellular* homology, used below). Thus, although different choices of the group of chains lead to different homology theories, in the case of a smooth manifold M, all homology theories provide isomorphic homology groups $H_k(M)$.

We shall not prove these unicity theorems nor shall present a complete construction of none of the theories (mentioning only the basics of cellular chains). Homology groups verify a number of simple properties proven in each theory, and hence there exist only *one* object, $\{H_k(\cdot)\}$, verifying all these properties. We describe below five such properties, hoping that the reader will know, understand and (mainly) be able to use them.

11.2.1 Invariance under homotopy (Properties 1 and 2)

We start with the simplest property of any homology theory:

Property 1 (*homology of a point*). *The homology groups of a point are $H_0(\text{pt}) = \mathbb{Z}$ and $H_k(\text{pt}) = 0$ for $k > 0$.*

The reader can prove it by taking the trivial triangulation. For the next statement see pp. 424-425.

Property 2 (*invariance under homotopy*). *Two homotopically equivalent manifolds or topological cell spaces* (defined below) X *and* X' *have isomorphic homology groups* $H_k(X) \approx H_k(X')$, $k \in \mathbb{N}$.

So, homeomorphic manifolds have isomorphic homology groups, be these manifolds diffeomorphic or not. Namely, the homology groups of all the 28 Milnor spheres (p. 110-111) are isomorphic to those of the usual sphere $\mathbb{S}^7$.

Corollary. *The homology groups of a topological space or manifold* X *homotopically equivalent to a point are* $H_0(\mathrm{X}) = \mathbb{Z}$ *and* $H_k(\mathrm{X}) = 0$ *for* $k > 0$.

Problem. Calculate the homology groups of the sphere $\mathbb{S}^n$, $n \geqslant 0$.

Answer. $H_0(\mathbb{S}^n, \mathbb{Z}) \approx H_n(\mathbb{S}^n, \mathbb{Z}) \approx \mathbb{Z}$ and $H_k(\mathbb{S}^n, \mathbb{Z}) = 0$ for $0 \neq k \neq n$.

Hint. A convenient triangulation of $\mathbb{S}^n$ (very useful in algebra!) is performed by the faces of all the dimensions $(0, \ldots, n)$ of the simplex Δ^{n+1} of dimension $n+1$. However, the resulting boundary operators (forming the so-called "Kozul complex" of the algebraists) are rather complicated, and one must take care on the orderings and indices in formula (1) of p. 410, to use correctly the orientations in the calculations (for instance, for $\mathbb{S}^{99}$ the simplex Δ^{100} has already lots of faces).

A different trick is to modify this triangulation of the sphere $\mathbb{S}^n$ as follows: the complement to the interior A of one of the n-dimensional faces is replaced by (or collapsed to) one chosen point $* \in \mathbb{S}^n$. This provides a generalised "triangulation" consisting of the point $*$ and of one set homeomorphic to an open n-dimensional ball, A.

The chain groups of this generalised "triangulation" are trivial, except $C_0 = \mathbb{Z}$ (generated by $*$) and $C_n = \mathbb{Z}$ (generated by A), giving the complex

$$0 \longrightarrow \mathbb{Z} \xrightarrow{\partial_n} 0 \longrightarrow \cdots 0 \xrightarrow{\partial_1} \mathbb{Z} \xrightarrow{\partial_0} 0.$$

In this case the boundary operators are trivial and the calculations are easy, provided that one understands why such a degenerate "triangulation" provides the same homology groups as any other – in p. 426, we shall explain it, passing from the true triangulation to the degenerate one.

Another justification of this so simple (but degenerate) "triangulation" is provided by cellular homology, without passing through a true triangulation. This "triangulation" is a "cellular decomposition" of $\mathbb{S}^n$ (see below).

11.2.2 Cell spaces and elements of cellular homology

A k-dimensional *cell* is a topological space homeomorphic to an open k-dimensional ball.

Manifolds and most topological spaces arising in geometry and physics are "cell spaces". Roughly speaking, this means that they are constructed by gluing together cells. For example, for any $k \geqslant 1$ the circle can be represented as the disjoint union of k 1-dimensional cells (open circle arcs) and k 0-dimensional cells (points); namely, one gets k 1-cells from a punctured circle at k points. Similarly, the "eight curve" ∞ (which is not a manifold) is the disjoint union of two 1-cells and one 0-cell. The sphere $\mathbb{S}^2$ can be constructed with one 2-cell and one 0-cell. More formally:

A *cell structure* on a topological space X is a decomposition of X into a disjoint union of cells σ_α^k (possibly of different dimensions). It is assumed that each homeomorphism $D^k \to \sigma_\alpha^k \subset X$ of the open disk D^k admits an extension to a continuous map (*characteristic map*) $\varphi_\alpha^k : \overline{D^k} \to X$ of the closed disk $\overline{D^k}$ (not necessarily injective on the boundary $\partial\overline{D^k}$), satisfying the following two conditions:

(i) The image of the boundary sphere $\partial\overline{D^k}$ is contained in a finite set of cells $\sigma_\beta^j \subset X$ of smaller dimension, $j < k$.

(ii) A subset $Y \subset X$ is closed if its intersection with the closure of any cell (of that cell structure) is closed.

Example. A k-dimensional convex polyhedron $P \subset \mathbb{R}^k$ (for instance, a simplex) has a natural cell structure. Namely, the interior of P is a k-dimensional cell, and the interior of each j-dimensional face is a j-cell, $0 \leqslant j < k$.

For the purpose of cellular homology each cell of a cell space must be oriented. The *group* $C_k(X)$ *of cellular* k*-dimensional chains* of a cell space X (for instance, a manifold) is the free commutative group that consist of all finite sums of k-dimensional cells with integral multiplicities.

The *boundary homomorphism* $\partial_k : C_k(X) \to C_{k-1}(X)$ sends each k-dimensional oriented cell σ_α^k to the $(k-1)$-dimensional cellular chain

$$\partial\sigma_\alpha^k = \sum_\beta [\sigma_\alpha^k : \sigma_\beta^{k-1}]\sigma_\beta^{k-1}\,,$$

where the integer $[\sigma_\alpha^k : \sigma_\beta^{k-1}]$, called *incidence coefficient* of σ_β^{k-1} in σ_α^k, is the multiplicity with which the oriented cell σ_β^{k-1} enters the boundary of σ_α^k.

It is the algebraic number of times that the oriented sphere $\partial\overline{D^k}$ covers the oriented cell σ_β^{k-1} under the characteristic map of σ_α^k, $\varphi_\alpha^k : \overline{D^k} \to X$.

Example. Both the torus $\mathbb{T}^2$ and the Klein bottle B^2 have a cell structure with one 2-cell, two 1-cells and one 0-cell. In fact, the representations of the torus and of the Klein bottle as a closed square (homeomorphic to $\overline{D^2}$) in which the opposite sides are identified with prescribed orientations, are two explicit descriptions of the respective characteristic maps $\varphi_T^2 : \overline{D^2} \to \mathbb{T}^2$ and $\varphi_B^2 : \overline{D^2} \to B$ of the 2-cell σ^2 of $\mathbb{T}^2$ and of B^2.

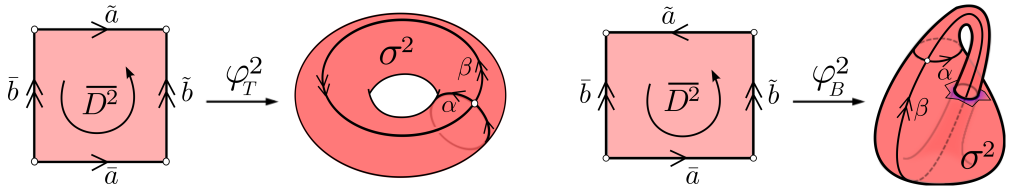


Figure 11.7: Characteristic maps of the 2-cell of the torus and the Klein bottle.

In Fig. 11.7, for both surfaces we have $\varphi_*^2(\tilde{a}) = \varphi_*^2(\bar{a}) = \alpha$ and $\varphi_*^2(\tilde{b}) = \varphi_*^2(\bar{b}) = \beta$ and $[\sigma^2 : \beta] = 0$ because $\tilde{b}$ and $\bar{b}$ have opposite orientations in $\partial\overline{D^2}$. However, the orientations of $\tilde{a}$ and $\bar{a}$ in $\partial\overline{D^2}$ provide the incidence coefficients $[\sigma^2 : \alpha] = 0$ for the torus (thus $\partial\sigma^2 = 0$) and $[\sigma^2 : \alpha] = 2$ for the Klein bottle (thus $\partial\sigma^2 = 2\alpha$). The reader can verify that this leads to $H_2(B^2, \mathbb{Z}) = 0$ and $H_1(B^2, \mathbb{Z}) = \mathbb{Z}_2 \times \mathbb{Z}$.

To give a rigorous definition of the incidence coefficient, we denote by $\mathrm{sk}^m(X)$ the union of cells of X whose dimension does not exceed $m \in \mathbb{N}$.

On the one hand, with each $(k-1)$-dimensional cell σ_β^{k-1} are naturally associated the sphere $\mathbb{S}_\beta^{k-1} := \mathrm{sk}^{k-1}(X)/(\mathrm{sk}^{k-1}(X) \setminus \sigma_\beta^{k-1})$ (all whose points but one belong to σ_β^{k-1}) and the trivial projection $\pi_\beta^{k-1} : \mathrm{sk}^{k-1}(X) \to \mathbb{S}_\beta^{k-1}$.

On the other hand, with each k-dimensional cell σ_α^k is naturally associated a map $\psi_\alpha^k : \partial\overline{D^k} \to \mathrm{sk}^{k-1}(X)$, which is the restriction of its characteristic map, φ_α^k, to the oriented boundary sphere $\partial\overline{D^k} (= \mathbb{S}^{k-1})$.

Thus, with a couple of cells σ_α^k and σ_β^{k-1} is naturally associated the composition $\pi_\beta^{k-1} \circ \psi_\alpha^k$ of their associated maps. Their incidence coefficient is defined as the degree of this map, $[\sigma_\alpha^k : \sigma_\beta^{k-1}] := \deg(\pi_\beta^{k-1} \circ \psi_\alpha^k : \partial\overline{D^k} \to \mathbb{S}_\beta^{k-1})$.

Example (Cells in triangulations). A triangulation of a manifold M (or of a topological space) provides a cell space structure on M for which the cellular chain groups are naturally isomorphic to the simplicial ones (replacing closed faces by the open ones). Clearly, the corresponding cellular and simplicial boundary operators coincide under this isomorphism, providing identical chain complexes and then isomorphic homology groups.

The semi-exactness condition $\partial^2 = 0$, necessary to have a chain complex, is satisfied, but we will not prove it. Cellular homology is very effective for concrete calculations.

11.2.3 Betti numbers and Euler characteristic

If the space of coefficients used to construct the chains is a field, then the homology "groups" are not simply groups, but they have also a structure of vector space over that field.

The kth *Betti number*[*] of a manifold M, is the dimension b_k of the vector space

$$H_k(M, \mathbb{R}) \approx \mathbb{R}^{b_k}.$$

Remark. The Betti numbers are defined in the same way for any field, not only for $\mathbb{R}$. The values of the Betti numbers depend on the chosen field.
An interesting fact (that we will not prove) is that for all the fields F of characteristic zero ($\mathbb{R}$, $\mathbb{C}$, or $\mathbb{Q}$) the dimension b_k of the vector space $H_k(M, F)$, that is, the k-th Betti number, coincides with the rank of the free part of the integral homology,

$$H_k(M, \mathbb{Z}) \simeq \mathbb{Z}^{b_k} \oplus \bigoplus_j \mathbb{Z}_{k_j}, \qquad b_k = \dim_F \, H_k(M, F).$$

Thus, in general, if the integral homology has torsion, it provides more information than the real one.
Moreover, the Betti numbers for the $\mathbb{Z}_2$ coefficients are often bigger than for the $\mathbb{R}$ coefficients (if there is even torsion in the integral homology). Hence, some times it is convenient to use the field $\mathbb{Z}_2$, since the calculations are easier and often one gets more information than using the field $\mathbb{R}$.

The *Euler characteristic* of an n-dimensional cell space X is the alternated sum of the numbers $|C_k|$ of k-dimensional cells $\chi(X) = \sum_{k=0}^{n}(-1)^k|C_k|$.

For example, for a Polyhedron, a polytope or a triangulated manifold it is the alternated sum of the numbers $|C_k|$ of k-dimensional faces.

The Euler characteristic of a manifold (or of a cell space) X is independent of any particular cell decomposition, polyhedral realisation or triangulation of X. It depends only on the "topological nature" of X.

Problem. Prove that the Euler characteristic of a cell space X is equal to the alternate sum of the Betti numbers $\chi(X) = \sum_{k=0}^{n}(-1)^k b_k$.
Hint: By definition $b_k = \dim\left(H^k(X, \mathbb{R})\right) = \dim\left(\operatorname{Ker}\partial_k\right) - \dim\left(\operatorname{Im}\partial_{k+1}\right)$.

[*] These numbers were introduced by Poincaré, rather than by E. Betti whose name Poincaré gave to these numbers.

Hence, *if two cell spaces have the same homotopy type, then they have also the same Euler characteristic.*

The Betti numbers depend on the field used to calculate them (see the above remark), but the Euler characteristic χ is independent of that field.

In the problems below the reader may use convenient triangulations or may invent other tricks – some of these tricks will be discussed later.

Problem. Compute the Betti numbers $b_0, b_1, \dots, b_n$ of the n-dimensional torus $\mathbb{T}^n = (\mathbb{S}^1)^n$.

Answer. These Betti numbers are the binomial coefficients $b_k(\mathbb{T}^n) = C_n^k$. For example, for $n = 3$ the integer homologies are

$$H_0(\mathbb{T}^3, \mathbb{Z}) \approx \mathbb{Z}, \quad H_1(\mathbb{T}^3, \mathbb{Z}) \approx \mathbb{Z}^3, \quad H_2(\mathbb{T}^3, \mathbb{Z}) \approx \mathbb{Z}^3, \quad H_3(\mathbb{T}^3, \mathbb{Z}) \approx \mathbb{Z}.$$

To get them, one can choose the circle-factors (a, b, c) as generating cycles for H_1, their pairwise products $a \times b$, $b \times c$, $c \times a$ as generating cycles for for H_2, and the fundamental cycle $a \times b \times c$ for H_3. Then, using the above remark, one obtains the Betti numbers $b_0 = 1$ $b_1 = 3$, $b_2 = 3$, $b_3 = 1$.

Remark that interpreting a circle as a cycle we represent it as the formal sum of the segments forming this circle, for any subdivision of it into segments (treating these segments as one-dimensional simplices).

Problem. Compute the homology groups and the Betti numbers of a compact connected closed orientable surface M^2 of genus g.

Answer. $H_0(M^2, \mathbb{Z}) \approx \mathbb{Z}$, $H_1(M^2, \mathbb{Z}) \approx \mathbb{Z}^{2g}$, $H_2(M^2, \mathbb{Z}) \approx \mathbb{Z}$.

Hint. See Ch. 5, p. 165 where we describe a convenient triangulation of the surface of genus $g = 2$. This surface may be described as the connected sum of two torus surfaces joined by a small cylindrical tube, providing in this way the triangulation generated by the triangulations of the torus summands.

Problem. Compute the homology groups, Betti numbers and Euler characteristic of the projective plane $\mathbb{R}\mathrm{P}^2$ for the coefficient groups $\mathbb{Z}$ and $\mathbb{Z}_2$.

Answer. $H_0(\mathbb{R}\mathrm{P}^2, \mathbb{Z}) = \mathbb{Z}$, $H_1(\mathbb{R}\mathrm{P}^2, \mathbb{Z}) = \mathbb{Z}_2$, $H_2(\mathbb{R}\mathrm{P}^2, \mathbb{Z}) = 0$;
$H_0(\mathbb{R}\mathrm{P}^2, \mathbb{Z}_2) = \mathbb{Z}_2$, $H_1(\mathbb{R}\mathrm{P}^2, \mathbb{Z}_2) = \mathbb{Z}_2$, $H_2(\mathbb{R}\mathrm{P}^2, \mathbb{Z}_2) = \mathbb{Z}_2$.
Thus the $\mathbb{R}$-Betti numbers are $b_0 = 1$, $b_1 = b_2 = 0$ while the $\mathbb{Z}_2$-Betti numbers are $b_0 = b_1 = b_2 = 1$, but both Euler characteristics coincide:

$$\chi(\mathbb{R}\mathrm{P}^2)_{\mathbb{R}} = 1 - 0 + 0 = 1 \quad \text{and} \quad \chi(\mathbb{R}\mathrm{P}^2)_{\mathbb{Z}_2} = 1 - 1 + 1 = 1.$$

Hint. The projective plane can be represented as the disjoint union of three cells of dimensions 0, 1 and 2: $\mathbb{R}\mathrm{P}^0 \subset \mathbb{R}\mathrm{P}^1$ and $\mathbb{R}^2 = \mathbb{R}\mathrm{P}^2 \setminus \mathbb{R}\mathrm{P}^1$. Thus, it suffices to consider the complex formed by the chains constructed from these three elements, with coefficients belonging to $\mathbb{Z}$ or to $\mathbb{Z}_2$.

The fact that the complement to the disc in $\mathbb{R}\mathrm{P}^2$ is the Möbius band, provides the relation $\partial A = 2a$ for the 2-dimensional and 1-dimensional generators A and a. The 1-boundaries are therefore the even multiples of a. The quotient group of the integers by the subgroup of the even integers provides $H_1(\mathbb{R}\mathrm{P}^2, \mathbb{Z}) \approx \mathbb{Z}/(2\mathbb{Z}) = \mathbb{Z}_2$. The same relation $\partial A = 2a$ implies the triviality of the group of 2-cycles, $Z_2(\mathbb{R}\mathrm{P}^2, \mathbb{Z}) = 0$, whence $H_2(\mathbb{R}\mathrm{P}^2, \mathbb{Z}) = 0$.

Similarly, $\partial A = 2a$ implies $H_1(\mathbb{R}\mathrm{P}^2, \mathbb{Z}_2) = \mathbb{Z}_2$, $H_2(\mathbb{R}\mathrm{P}^2, \mathbb{Z}_2) = \mathbb{Z}_2$.

Problem. Calculate the homology groups and the Betti numbers of the real projective space $\mathbb{R}\mathrm{P}^n$.

Answer. $H_{2k-1}(\mathbb{R}\mathrm{P}^n, \mathbb{Z}) = \mathbb{Z}_2$ for $2k-1 < n$, $H_n(\mathbb{R}\mathrm{P}^n, \mathbb{Z}) = \mathbb{Z}$, for odd n, and $H_{2k}(\mathbb{R}\mathrm{P}^n, \mathbb{Z}) = 0$ for $k > 0$.

Example. For $\mathbb{R}\mathrm{P}^3 \sim \mathrm{SO}(3)$ the Betti numbers are ($b_0 = 1$, $b_1 = 0$, $b_2 = 0$, $b_3 = 1$). The homology group $H_1(\mathbb{R}\mathrm{P}^3, \mathbb{Z}) \approx \mathbb{Z}_2$ is non trivial.

The generators of the homology groups $H_{2k-1}(\mathbb{R}\mathrm{P}^n, \mathbb{Z}) = \mathbb{Z}_2$ are the fundamental classes of the (oriented) closed submanifolds $\mathbb{R}\mathrm{P}^{2k-1}$.

The homology group $H_{2k-1}(\mathbb{R}\mathrm{P}^{2k-1}, \mathbb{Z}) = \mathbb{Z}$ is generated by the fundamental class of the orientable manifold $\mathbb{R}\mathrm{P}^{2k-1}$.

Problem. Calculate the homology groups of the complement U to n different points of the plane $\mathbb{R}^2$.

Answer. $H_0(U, \mathbb{Z}) = \mathbb{Z}$, $H_1(U, \mathbb{Z}) = \mathbb{Z}^n$, $H_{k>1}(U, \mathbb{Z}) = 0$.

Hint. To compute the homology of this non-compact manifold, use the homotopy invariance of homology (Property 2, p. 415). For $n = 1$ the manifold U can be retracted to the circle $\mathbb{S}^1$, and this retraction provides an (easy) proof of the natural isomorphisms

$$H_*(U, \mathbb{Z}) \approx H_*(\mathbb{S}^1, \mathbb{Z}).$$

(A retraction of a space X to its subspace Y is a continuous family of maps $f_t : X \to X$, $t \in [0,1]$, such that f_0 is the identity map, the restriction of f_t to Y is the identity map for all t, and such that $f_1(X) \subset Y$.)

Similarly, for $n = 2$ the domain U is continuously retractable to the "∞" curve, and for any n to the "bouquet of n circles": Obtained from n disjoint circles $\mathbb{S}^1_k$ by gluing a point $x_k \in \mathbb{S}^1_k$ (for $k = 1, \dots, n$) to a common point $*$.

To compute the homologies of the resulting curve, it suffices to consider the complex generated by the point $*$ as the only 0-cell, and by the n 1-cells $\mathbb{S}^1_k \setminus \{x_k\}$. The circles $\mathbb{S}^1_k$ are n independent cycles generating $H_1(U, \mathbb{Z})$.

Problem. Compute the homology groups of the projective plane $\mathbb{C}\mathrm{P}^2$.

Answer. $H_0(\mathbb{C}\mathrm{P}^2, \mathbb{Z}) = \mathbb{Z}$, $H_1(\mathbb{C}\mathrm{P}^2, \mathbb{Z}) = 0$, $H_2(\mathbb{C}\mathrm{P}^2, \mathbb{Z}) = \mathbb{Z}$, $H_3(\mathbb{C}\mathrm{P}^2, \mathbb{Z}) = 0$, $H_4(\mathbb{C}\mathrm{P}^2, \mathbb{Z}) = \mathbb{Z}$.

Hint. Represent the projective plane $\mathbb{C}\mathrm{P}^2$ as the union of a point, that is, the 0-cell $(\infty = \mathbb{C}\mathrm{P}^0) \in \mathbb{C}\mathrm{P}^1$, with the 2-cell $\mathbb{C}\mathrm{P}^1 \setminus \mathbb{C}\mathrm{P}^0 \approx \mathbb{C}$, and of the remaining 4-cell $\mathbb{C}\mathrm{P}^2 \setminus \mathbb{C}\mathrm{P}^1 \approx \mathbb{C}^2$.

The resulting complex has no odd dimensional chains, and provides the above homology groups.

Problem. Find the homology class of a smooth complex algebraic curve X in the projective plane $\mathbb{C}\mathrm{P}^2$.

Answer. The curve belongs to the homology class of the cycle $\mathbb{C}\mathrm{P}^1$, multiplied by the degree n of the curve:

$$[X] = n\,[\mathbb{C}\mathrm{P}^1].$$

Solution. Since the group $H_2(\mathbb{C}\mathrm{P}^2) = \mathbb{Z}$ is generated by the cycle $\mathbb{C}\mathrm{P}^1$, the homology class of any 2-cycle is an integral multiple of $[\mathbb{C}\mathrm{P}^1]$. Thus it remains to show that this integer is n.

First, take the degenerate curve $\widetilde{X}$ of degree n formed by n straight lines. A homology between $\widetilde{X}$ and $n\,[\mathbb{C}\mathrm{P}^1]$ can be easily constructed explicitly.

Next, consider the curve $\widehat{X}$ defined by the equation $f = \varepsilon$, where $f = 0$ is the equation of the curve $\widetilde{X}$. Its homology to $\widetilde{X}$ is provided by the family of curves defined by the equations $f = \varepsilon t$, $0 \leqslant t \leqslant 1$.

Finally, join the non degenerate curves $\widehat{X}$ and X by a smooth path in the space of the non degenerate curves of degree n. Such path exists, by the "Italian principle", since the real codimension of the set of degenerate curves of degree n is at least 2 in the space of curves of degree n. This path provides a homology between X and $\widehat{X}$, and hence $[X] = [\widehat{X}] = [\widetilde{X}] = n\,[\mathbb{C}\mathrm{P}^1]$.

Remark. The algebraic curve X is a smooth compact closed oriented real surface of genus $g = \frac{(n-1)(n-2)}{2}$ (Riemann-Hurwitz formula of p. 149). R. Thom conjectured many years ago that the homology class $n\,[\mathbb{C}\mathrm{P}^1] \in H_2(\mathbb{C}\mathrm{P}^2, \mathbb{Z})$ cannot be represented by a smooth real surface of smaller genus than the genus $g = (n-1)(n-2)/2$ of the complex smooth plane algebraic curves of degree n.

This conjecture was proved by Kronheimer. The proof being difficult and depending on ideas and technology of quantum field theory, we shall not discuss it in this book.

Cobordisms. The k-homology classes in a higher dimensional manifold M are not always representable by the smooth submanifolds of dimension k in M: our chains may have complicated singularities.

A classification of the smooth submanifolds of dimension k in a manifold M, called *cobordism* of that submanifolds in M, starts with the relation $X \sim X'$ defined as the existence of a smooth submanifold σ of dimension $k+1$ whose boundary is their difference,

$$\partial\sigma^{k+1} = X - X'.$$

This classification is different from the homologies between the k-chains X and X' (for a homology the chain σ^{k+1} might be very singular).

To make the cobordism an equivalence relation, one says that two k-manifolds X and X' of M are *cobordant* if there exists an oriented submanifold σ^{k+1} (called *cobordism* between X and X') of the product space $M \times [0,1]$ such that $\partial\sigma^{n+1} = X \times \{1\} - X' \times \{0\}$. Cobordant oriented submanifolds are homologous, but generically there are many cobordism classes in a given homology class.

11.2.4 Induced homomorphisms (Property 3)

If two k-chains X and X' are homotopic, they are homologous. The connecting "cylindrical" $(k+1)$-chain is provided by the homotopy, as it is explained in Ch. 7, p. 237, where we have applied this homotopy of chains to the Lie derivative calculation.

However, homologous chains are sometimes represented by non homotopic embedded submanifolds. For instance, the small circle cutting the connecting tube of the connected sum M of two tori, is homologous to zero on the surface of genus $g = 2$, which is the connected sum: This circle is the boundary of the part provided by one of the two tori. However, this embedded circle is not homotopic to zero: It does not bound any disc-chain in M.

In spite of this difference, the relations between homotopy and homology are numerous and extremely useful for both theories.

A continuous map $f : M \to \widetilde{M}$ between two manifolds (or spaces) M and $\widetilde{M}$ induces the natural homomorphisms

$$f_* : H_k(M) \to H_k(\widetilde{M})$$

provided by the "natural" homomorphisms f_{*k} of the corresponding chain groups of the complexes

$$\begin{tikzcd}
\cdots \arrow[r] & C_k(M) \arrow[r, "\partial_k"] \arrow[d, "f_{*k}"'] & C_{k-1}(M) \arrow[r] \arrow[d, "f_{*k-1}"] & \cdots \\
\cdots \arrow[r] & C_k(\widetilde{M}) \arrow[r, "\widetilde{\partial}_k"] & C_{k-1}(\widetilde{M}) \arrow[r] & \cdots
\end{tikzcd}$$

To explain this, one has to keep in mind that the image of a cycle is a cycle whatever be the definition of the chains. In particular, the image of the boundary of a chain is the boundary of the image of that chain, that is, $\widetilde{\partial}_k(f_{*k}c) = f_{*k-1}(\partial_k c)$.

But the way in which f_* is defined depends on the choice of the theory. Namely, we need to arrange that f_* sends the chains of M to the chains of $\widetilde{M}$. This is naturally done for the singular chains.

The construction of these induced homomorphisms of chain complexes extends naturally to the composition of maps.

Property 3 (*functoriality*). *The composition $f\circ g$ of any pair of continuous maps $g : M \to \widetilde{M}$ and $f : \widetilde{M} \to \widehat{M}$ satisfies $(f\circ g)_* = f_*\circ g_*$.*

the homology groups are discrete invariants...

The isomorphism of the homology groups of homotopy equivalent topological spaces (Property 2 above) follows from the invariance property.

These induced homomorphisms f_{*k} are invariant under homotopy.

Property $\tilde{\mathbf{2}}$ (*homotopic invariance*). *If two continuous maps $f, g : M \to \widetilde{M}$ are homotopic, then their induced homomorphisms coincide: $f_* = g_*$.*

Definition. A continuous map $f : M \to \widetilde{M}$ is called a *homotopy equivalence* if there exists a continuous map $g : \widetilde{M} \to M$ with the following properties:

1) The product $g\circ f : M \to M$ is homotopic to the identity map of M.

2) The product $f\circ g : \widetilde{M} \to \widetilde{M}$ is homotopic to the identity map of $\widetilde{M}$.

Theorem 1. *If a map f is a homotopy equivalence, then its induced homomorphisms f_{*k} are isomorphisms of the homology groups.*

Proof. The product $g_{*k} \circ f_{*k} : H_k(M) \to H_k(M)$ is the identity map because $g \circ f$ is homotopic to the identity, and a homotopy is unable to change the homology classes.

Similarly, the product $f_{*k} \circ g_{*k} : H_k(\widetilde{M}) \to H_k(\widetilde{M})$ is the identity map.

In consequence, the two induced maps

$$f_{*k} : H_k(M) \to H_k(\widetilde{M}), \quad g_{*k} : H_k(\widetilde{M}) \to H_k(M)$$

are isomorphisms (inverse to each other). □

Example. Consider the inclusion $i : M \to N$ of the lemniscate curve "∞", noted by M, into the complement of the two central points of the lemniscate loops in the plane, $N = \mathbb{R}^2 \setminus \mathbb{S}^0$.

Consider the family $\{j_t : N \to N\}$ of continuous contractions of N onto the neighbourhoods $j_t(N)$ of the lemniscate M, which leaves each point of the lemniscate fixed for all t in the interval $0 \leqslant t \leqslant 1$, which sends N to the lemniscate M for $t = 1$ and which is the identity map of N for $t = 0$.

An explicit construction of such a family, called *retraction* of N to M, is not difficult: One decomposes N into the trajectories of the homotopy j_t, which lead to the different points of the lemniscate.

The product $j_1 \circ i : M \to M$ is the identity map of M. The homotopy of the product $i \circ j_1 : N \to N$ to the identity map of N is provided by the family j_t of maps $N \to N$ (where j_0 is the identity map of N and j_1 is $i \circ j_1$).

Therefore, the inclusion $i : M \to N$ is a homotopy equivalence, and we obtain the isomorphisms (used above in the hint of page 420):

$$i_{*k} : H_k(M) \to H_k(N).$$

11.2.5 Relative homologies (Properties 4 and 5)

Another important general property of the homology groups is their behaviour for the pairs of manifolds (or spaces) $Y \subset X$.

Definition. The *relative k-chains* of X modulo Y are defined as the elements of the Abelian group

$$C_k(X, Y) = \frac{C_k(X)}{i_* C_k(Y)},$$

where $i_* : C_k(Y) \to C_k(X)$ is the map of chains, induced by the inclusion $i : Y \to X$ of the submanifold (subspace) Y into the manifold (space) X.

Since the natural maps i_* commute with the boundary operators, we naturally define the group of *relative k-cycles* $Z_k(X, Y)$, which consist of those k-chains of X, whose boundaries belong to $i_* C_{k-1}(Y)$, and also the group of relative k-boundaries $B_k(X, Y)$:

A k-chain c in X is a *relative k-boundary* if there exist a $(k+1)$-chain C in X, whose boundary $\partial C = c + \delta$ differs from c by a k-chain δ belonging to the group $i_* C_k(Y)$.

The relative k-boundaries are relative k-cycles, and hence the relative chains form a chain complex. Its homologies are called *relative homologies* of X modulo Y,

$$H_k(X, Y) = Z_k(X, Y)/B_k(X, Y).$$

Essentially, this construction provides the homology of the space X/Y, which is the space X in which the subspace Y is replaced by one point $*$, $X \smallsetminus Y = (X/Y) \smallsetminus *$. Namely, the following isomorphisms hold:

$$H_k(X, Y) \approx H_k(X/Y, *) \approx \begin{cases} H_k(X/Y), & k > 0, \\ H_0(X/Y)/\mathbb{Z}, & k = 0. \end{cases}$$

To avoid the special account for $k = 0$, the latter groups are called *reduced homology groups* of the space X/Y and denoted by $\overline{H}_k(X/Y)$, that is, $\overline{H}_{k>0}(X/Y) := H_{k>0}(X/Y)$ and $\overline{H}_0(X/Y) := H_0(X/Y)/\mathbb{Z}$. Summarising:

Property 4 (*relative-quotient*). $H_k(X, Y) \approx H_k(X/Y, *) \approx \overline{H}_k(X/Y)$.

In general, the space X/Y is not a smooth manifold, even when Y is a smooth submanifold of X. So, to avoid the difficulties of the formal description of the k-chains of a bad space X/Y, it suffices to replace the ordinary chains by the relative chains.

We can do it even in the case when neither X nor Y are smooth manifolds, for instance, for the case of singular algebraic varieties: The only condition needed here, is that i_* sends naturally the chains of Y to the chains of X (whatever be the definition of the chains).

The main property of the relative homologies is

Property 5 (*long exact sequence*). *The following exact sequence holds*:

$$\ldots \to H_k(Y) \xrightarrow{i_*} H_k(X) \xrightarrow{p_*} H_k(X/Y) \xrightarrow{D} H_{k-1}(Y) \to \ldots \tag{2}$$

Here, the natural map

$$p_* : H_k(X) \to H_k(X/Y)$$

models the projection $p : (X, Y) \to (X/Y, *)$, which sends Y to the point $*$ of the "quotient space" X/Y, and is the identity map of the complement, $(X \setminus Y) \to ((X/Y) \setminus *)$.

The most important is the "joining homomorphism"

$$D : H_k(X/Y) \to H_{k-1}(Y)\,,$$

which is also quite natural: A relative homology class $[c] \in H_k(X/Y)$ is the homology class of a relative k-cycle, that is, the class of a chain $c \in C_k(X)$ whose boundary $\partial c \in C_{k-1}(X)$ "belongs" to Y (formally, it is the image of a $(k-1)$-chain of Y by the inclusion homomorphism i_*). This $(k-1)$-chain of Y is a cycle because $\partial^2 c = 0$, and its homology class is the required element $D[c] \in H_{k-1}(Y)$.

The independence of $D[c]$ from the choice of the representative c of $[c]$ can be easily proved. The exactness of the long sequence (2) is also an elementary fact. We can use this fact to calculate the homology groups of the previous examples, like those of the manifolds $\mathbb{S}^n$, $\mathbb{R}\mathrm{P}^n$ and $\mathbb{C}\mathrm{P}^n$.

For instance, for the natural triangulation of the boundary surface X of the n-simplex (say, of the tetrahedron surface decomposed into 4 vertices, 6 edges and 4 faces), we can consider as Y the part formed by the complement to one of the faces. This part Y is contractible and thus has trivial homologies ($H_k(Y) = 0$ for $k > 0$). Therefore, the exact sequence (2) provides the isomorphisms $p_* : H_k(X) \to H_k(X/Y)$. We have thus reduced the calculations of the homologies of the big chain complex of the triangulation X, to those of the trivial complex of the relative chains $C_k(X/Y)$, which provides isomorphic homologies. This is the geometric foundation of the calculations of the homologies of the spheres, done in page 415.

A different trick is to consider the sphere $\mathbb{S}^n$ as a simplex X of dimension n whose boundary Y (formed by lower dimensional simplices) is collapsed to a chosen point $* \in \mathbb{S}^n$, providing again the generalised "triangulation" consisting of the point $*$ and of one n-dimensional cell.

Exercise. Use this trick to compute the homology groups of the sphere $\mathbb{S}^n = X/Y$ by means the exact sequence of the pair $Y \subset X$.

11.2.6 Some explicit computations of homology groups

Problem. Compute the homology groups of the projective space $\mathbb{C}\mathrm{P}^n$.

Answer. $H_{2k}(\mathbb{C}\mathrm{P}^n, \mathbb{Z}) = \mathbb{Z}$ for $k \leqslant n$, $H_{2k-1}(\mathbb{C}\mathrm{P}^n, \mathbb{Z}) = 0$.

Solution. We shall prove it by induction, deducing the $\mathbb{C}\mathrm{P}^n$ case from the $\mathbb{C}\mathrm{P}^{n-1}$ case. For $n = 1$, where $\mathbb{C}\mathrm{P}^1 = \mathbb{S}^2$, the answer is evidently correct. For the submanifold $Y = \mathbb{C}\mathrm{P}^{n-1}$ naturally embedded in $X = \mathbb{C}\mathrm{P}^n$, the exact sequence (2) has the form

$$H_{k+1}(\mathbb{S}^{2n}, \mathbb{Z}) \xrightarrow{D} H_k(\mathbb{C}\mathrm{P}^{n-1}, \mathbb{Z}) \xrightarrow{i_*} H_k(\mathbb{C}\mathrm{P}^n, \mathbb{Z}) \xrightarrow{p_*} H_k(\mathbb{S}^{2n}, \mathbb{Z}),$$

since $X/Y = \mathbb{C}^n/\infty = \mathbb{S}^{2n}$.

If $k + 1 < 2n$, then the leftmost and the rightmost groups are trivial ($= 0$), and hence the homomorphism i_* is an isomorphism. Thus, we have proved the part of the answer corresponding to $k < 2n - 1$, for $H_*(\mathbb{C}\mathrm{P}^n)$.

For $k = 2n - 1$ the inductive assumption provides the exact sequence

$$(H_{2n-1}(\mathbb{C}\mathrm{P}^{n-1}) = 0) \xrightarrow{i_*} H_{2n-1}(\mathbb{C}\mathrm{P}^n) \xrightarrow{p_*} (H_{2n-1}(\mathbb{S}^{2n}) = 0),$$

and hence $H_{2n-1}(\mathbb{C}\mathrm{P}^n, \mathbb{Z}) = 0$.

For $k = 2n$ the exact sequence (2) provides the segment

$$(H_{2n}(\mathbb{C}\mathrm{P}^{n-1}) = 0) \xrightarrow{i_*} H_{2n}(\mathbb{C}\mathrm{P}^n) \xrightarrow{p_*} (H_{2n}(\mathbb{S}^{2n}) = \mathbb{Z}) \xrightarrow{D} H_{2n-1}(\mathbb{C}\mathrm{P}^{n-1}).$$

Since the rightmost group is trivial ($H_{2n-1}(\mathbb{C}\mathrm{P}^{n-1}) = 0$), we conclude that p_* is an isomorphism, and hence $H_{2n}(\mathbb{C}\mathrm{P}^n, \mathbb{Z}) = \mathbb{Z}$.

Finally, for $k > 2n = \dim \mathbb{C}\mathrm{P}^n$, we have evidently $H_k(\mathbb{C}\mathrm{P}^n) = 0$.

Problem. Calculate the relative homologies of the pair: $X = \mathbb{S}^1 \times B^2$, $Y = \mathbb{S}^1 \times (\partial B^2)$, where B^2 is the closed disc in $\mathbb{R}^2$.

Solution. Consider the Hopf fibration $\pi : \mathbb{S}^3 \to \mathbb{S}^2$ (with fibre $\mathbb{S}^1$) and the embedding of a closed hemisphere $B^2 \to \mathbb{S}^2$. Denote $\widetilde{X} = \mathbb{S}^3$ and write $\widetilde{Y}$ for the preimage $\pi^{-1}(B^2)$. Clearly, we have $\widetilde{X} \setminus \widetilde{Y} = X \setminus Y$ and $X/Y \approx \widetilde{X}/\widetilde{Y}$. The obvious retraction of $\widetilde{Y}$ to the fibre $\mathbb{S}^1$ of the Hopf fibration proves that $H_k(\widetilde{Y}) = H_k(\mathbb{S}^1)$.

Moreover, the exact sequence of the pair $Y \subset X$ is provided by the following sequence of homomorphisms for the pair $\widetilde{Y} \subset \widetilde{X}$,

$$H_3(\widetilde{Y}) \to H_3(\widetilde{X}) \to H_3(\widetilde{X}/\widetilde{Y}) \xrightarrow{D_3} H_2(\widetilde{Y}) \to H_2(\widetilde{X}) \to H_2(\widetilde{X}/\widetilde{Y}) \xrightarrow{D_2} H_1(\widetilde{Y}) \to$$

$$\to H_1(\widetilde{X}) \to H_1(\widetilde{X}/\widetilde{Y}) \xrightarrow{D_1} H_0(\widetilde{Y}) \to H_0(\widetilde{X}) \to \overline{H}_0(\widetilde{X}/\widetilde{Y}).$$

In this sequence we know the groups:

$$H_3(\widetilde{X}) = \mathbb{Z}, \quad H_2(\widetilde{X}) = 0, \quad H_1(\widetilde{X}) = 0, \quad H_0(\widetilde{X}) = \mathbb{Z},$$

$$H_0(\widetilde{Y}) = \mathbb{Z}, \quad \overline{H}_0(\widetilde{X}/\widetilde{Y}) = 0, \quad H_1(\widetilde{Y}) = \mathbb{Z}, \quad H_2(\widetilde{Y}) = 0, \quad H_3(\widetilde{Y}) = 0.$$

Thus the above sequence reduces to

$$0 \to \mathbb{Z} \to H_3(\widetilde{X}/\widetilde{Y}) \xrightarrow{D_3} 0 \to 0 \to H_2(\widetilde{X}/\widetilde{Y}) \xrightarrow{D_2} \mathbb{Z} \to$$

$$\to 0 \to H_1(\widetilde{X}/\widetilde{Y}) \xrightarrow{D_1} \mathbb{Z} \xrightarrow{\sim} \mathbb{Z} \to 0,$$

which means that $H_3(\widetilde{X}/\widetilde{Y}) = \mathbb{Z}$, $H_2(\widetilde{X}/\widetilde{Y}) = \mathbb{Z}$, $H_1(\widetilde{X}/\widetilde{Y}) = 0$.

This provides also the generating relative cycles: The fundamental cycle of the solid torus X/Y for H_3 and the cycle B_2 (mod ∂B_2) for H_2.

Problem. Compute the homology groups of the 3-manifold T_1M formed by the tangent vectors of length 1 of a closed orientable connected 2-dimensional surface M of genus g (we suppose a Riemannian metric is given on M).

Answer. If $g = 0$, then $M \approx \mathbb{S}^2$ and the 3-manifold T_1M is diffeomorphic to $\mathrm{SO}(3) \approx \mathbb{R}\mathrm{P}^3$ (see p. 80), so that

$$H_0(T_1\mathbb{S}^2) = \mathbb{Z}\,, \quad H_1(T_1\mathbb{S}^2) = \mathbb{Z}_2\,, \quad H_2(T_1\mathbb{S}^2) = 0\,, \quad H_3(T_1\mathbb{S}^2) = \mathbb{Z}\,.$$

If $g = 1$, the manifold T_1M is the 3-torus $\mathbb{T}^3$. Hence (see p. 419)

$$H_0(T_1\mathbb{T}^2) = \mathbb{Z}\,, \quad H_1(T_1\mathbb{T}^2) = \mathbb{Z}^3\,, \quad H_2(T_1\mathbb{T}^2) = \mathbb{Z}^3\,, \quad H_3(T_1\mathbb{T}^2) = \mathbb{Z}\,.$$

If $g > 1$, besides $H_0(T_1M) = \mathbb{Z}$ we have the following $\mathbb{Z}$-homology groups

$$H_1(T_1M) = \mathbb{Z}^{2g} + \mathbb{Z}_{2g-2}, \quad H_2(T_1M) = \mathbb{Z}^{2g}, \quad H_3(T_1M) = \mathbb{Z}. \tag{3}$$

Solution. It remains to prove the answer for $g > 1$.

The manifold T_1M is fibred over M into circular fibres

$$\mathbb{S}^1 \to T_1M \xrightarrow{p} M.$$

Let U be a small open disc around a point of the surface M and E^3 be the preimage under p of $M \setminus U$, $E^3 = p^{-1}(M \setminus U) \subset T_1M$.

To compute the homologies of $X = T_1M$ we need the exact sequence (2) of Property 5 and the relative homologies $H_k(X/Y)$, where $Y = E^3$.

There exists a smooth vector field v of vectors of length 1 on $M \setminus U$, since we may send all the zeros of a generic vector field on M inside U by a diffeomorphism. Hence, E^3 is diffeomorphic to the direct product $\mathbb{S}^1 \times (M \setminus U)$. Indeed, a trivial projection $E^3 \to \mathbb{S}^1$ is given by sending each unit tangent vector of $M \setminus U$ to its azimuth $\varphi \in \mathbb{S}^1$ measured from the direction of v (using the orientation of M).

The homologies of E^3 are now easily computable. Namely,

$$H_1(E^3, \mathbb{Z}) = \mathbb{Z}^{2g+1},$$

since there are $2g$ generators provided by the surface M of genus g, and one more generator by the $\mathbb{S}^1$-factor of the product E^3.

The second homology group

$$H_2(E^3, \mathbb{Z}) = \mathbb{Z}^{2g}$$

is generated by the (toric) products of the $2g$ cycles of $M \setminus U$ with the factor $\mathbb{S}^1$ of E^3. We have also, evidently,

$$H_0(E^3, \mathbb{Z}) = \mathbb{Z}, \qquad H_3(E^3, \mathbb{Z}) = 0.$$

The relative homologies $H_k(X/Y)$, where $X = T_1M$ and $Y = E^3$, are the homologies of the solid torus $\mathbb{S}^1 \times B^2$ modulo its boundary (where B^2 denotes the closure of the open disc U). According to page 427, these homologies are

$$H_3(X/Y) = \mathbb{Z}, \quad H_2(X/Y) = \mathbb{Z}, \quad H_1(X/Y) = 0, \quad \overline{H}_0(X/Y) = 0.$$

To prove (3) we substite the known groups and homomorphisms in the long exact sequence (2) of Property 5, which has the form

$$H_3(X) \to H_3(X/Y) \xrightarrow{D_3} H_2(Y) \to H_2(X) \to H_2(X/Y) \xrightarrow{D_2} H_1(Y) \to$$

$$\to H_1(X) \to H_1(X/Y) \xrightarrow{D_1} H_0(Y) \xrightarrow{1} H_0(X) \to (\overline{H}_0(X/Y) = 0)\,,$$

to get the exact sequence

$$\mathbb{Z} \xrightarrow{1} \mathbb{Z} \xrightarrow{D_3} \mathbb{Z}^{2g} \to H_2(X) \to \mathbb{Z} \xrightarrow{D_2} \mathbb{Z}^{2g+1} \to H_1(X) \to 0 \xrightarrow{D_1} \mathbb{Z} \xrightarrow{1} \mathbb{Z} \to 0.$$

The main point is to calculate the joining homomorphism

$$D_2 : (H_2(X, Y) = \mathbb{Z}) \to (H_1(Y) = \mathbb{Z}^{2g+1}).$$

The generator of the group $H_2(X, Y)$ (computed on p. 427) is represented by a "parallel" vector field on the closed disc B^2. The boundary γ of this generator is represented by that parallel vector field along the boundary ∂B^2 of the disc, and then it is a closed curve on the torus formed by all the unit tangent vectors of M at the points of the cycle ∂B^2.

To understand the position of this curve γ with respect to the $2g + 1$ generators of the group $H_1(Y)$, where $Y = (M \setminus U) \times \mathbb{S}^1$, we note that the projection of γ to the base surface $M \setminus U$ is homologous to 0, while the index of γ with respect to the vector field parallelising $M \setminus U$, with no zeros inside $M \setminus U$, is equal to $2-2g$, since the sum of the indices of a smooth vector field on M is equal to the Euler characteristic $2 - 2g$ of M (by the Euler-Poincaré theory – see Ch. 10, p. 374).

In consequence, the curve γ is homologous to $2 - 2g$ times the generator β of $H_1(Y)$, representing the fibre $\mathbb{S}^1$.

Thus, if $g > 1$, then the kernel of the map $D_2 : \mathbb{Z} \to \mathbb{Z}^{2g+1}$ is trivial, while its image is the subgroup generated by $(2g-2)\beta$, where β is one of the generators of $\mathbb{Z}^{2g+1}$. We conclude that, for $g > 1$,

$$H_1(X) = \mathbb{Z}^{2g+1}/(\operatorname{Im} D_2) = \mathbb{Z}^{2g} + \mathbb{Z}/(2g-2)\mathbb{Z} = \mathbb{Z}^{2g} + \mathbb{Z}_{2g-2}$$

The same formula, except for the last group, provides the correct answer $H_1(X) = \mathbb{Z}^3$ even in the exceptional case $g = 1$ in which $M = \mathbb{T}^2$.

To calculate $H_2(X)$, the known relations $D_3 = 0$ and $\operatorname{Ker} D_2 = 0$ provide the isomorphism

$$H_2(X) \approx H_2(Y) \approx \mathbb{Z}^{2g} \quad (\text{for } g \neq 1).$$

Remark. The geometric construction of these calculations has been reformulated as an abstract algebraic algorithm, similar to the algorithm of the multiplication of integers written in the decimal system.

This algorithm, called "spectral sequence", will be discussed below. While all the proofs are completely trivial, the real calculations are sometimes long, and any minor misprint destroys the result completely: one should take care of several indices in complicated notations.

11.3 "Spectral sequences" in singularity theory

The spectral sequence algorithm is essentially a Newton type version of the successive approximation algorithms in the calculus of power series. We shall show now its application to a computational problem of algebra, where the length of the computations is relatively small with respect to the typical problems in topology, like those of the calculation of the homologies of the groups $\mathrm{SO}(n)$ or $\mathrm{U}(n)$.

Homological equation. We shall discuss the attempt to solve the "homological equation" (discussed in pp. 540-541): We wish to know whether a given function α belongs to the gradient ideal of a given function F of variables $(x_1, \ldots, x_n)$, that is, we wish to solve in g_k the equation

$$\alpha(x_1, \ldots, x_n) = \sum_{k=1}^{n} g_k(x_1, \ldots, x_n) \frac{\partial F(x_1, \ldots, x_n)}{\partial x_k}. \qquad (1)$$

Notations. To simplify the presentation, we shall introduce the following spaces of functions and of vector fields with their notations.

Write $\mathfrak{m}$ for the "maximal ideal" of power series in $(x_1, \ldots, x_n)$, which is the space of the series vanishing at $x = 0$: The elements of $\mathfrak{m}$ are the formal series $f = a_1 x + a_2 x^2 + \ldots$ if $n = 1$ ("$f = O(x)$" in the notations of calculus).

We write $\mathfrak{m}^p$ for the "p-th power of the maximal ideal": For $n = 1$ it is the set of the series $f = a_p x^p + a_{p+1} x^{p+1} + \ldots$ (in calculus, the notation would be $f = O(|x|^p)$). In algebra, the monomials ax^u ($x \in \mathbb{C}^n$, $u \in \mathbb{Z}^n$, $x^u = x_1^{u_1} x_2^{u_2} \ldots x_n^{u_n}$) that generate these series f, are supposed to be of degree $| u |= u_1 + \ldots + u_n \geqslant p$.

The notation for the space of vector fields of the form

$$w = \sum_{k=1}^{n} g_k \frac{\partial}{\partial x_k}, \qquad g_k \in \mathfrak{m}^{q+1},$$

will be $\mathfrak{a}_q$. The character $\mathfrak{a}$ is taken from the word "(Lie) algebra", and the degree q is due to the "natural filtration" conditions

$$\mathfrak{a}_q \mathfrak{m}^p \subset \mathfrak{m}^{p+q}, \qquad \mathfrak{m}^p \mathfrak{a}_q \subset \mathfrak{a}_{p+q}.$$

Remark. We shall not use it directly, but the Poisson brackets of vector fields are also filtered naturally:

$$\{\mathfrak{a}_p, \mathfrak{a}_q\} \subset \mathfrak{a}_{p+q}.$$

Similarly to the Laurent monomials, the basic fields should be considered as having negative degrees:

$$\deg \frac{\partial}{\partial x} = \deg \frac{1}{x} = -1, \quad \text{if} \ \deg x = 1.$$

The algebra of the power series is filtered by the powers of its maximal ideal:

$$\mathbb{C}[[x]] \supset \mathfrak{m} \supset \mathfrak{m}^2 \supset \mathfrak{m}^3 \supset \dots, \qquad \mathfrak{m}^p \mathfrak{m}^q \subset \mathfrak{m}^{p+q}.$$

The spaces $M_p = \mathfrak{m}^p / \mathfrak{m}^{p+1}$ form the *adjoined graded algebra*

$$\mathbb{C}[x] = \mathbb{C} \oplus \mathbb{C}^n \oplus \mathbb{C}^{n(n+1)/2} \oplus \dots.$$

Of course, M_p is the vector space of the homogeneous polynomials of degree p in the variables $(x_1, \dots, x_n)$ with coefficients in $\mathbb{C}$. It is isomorphic to the space of truncated series in $\mathfrak{m}^p$ (to obtain the class of a series in $\mathfrak{m}^p$, one keeps only its term of degree p).

Similarly, the filtration in the space of vector fields determines the spaces $A_q = \mathfrak{a}_q / \mathfrak{a}_{q+1}$ and the associated graded vector space

$$A = A_{-1} \oplus A_0 \oplus A_1 \oplus A_2 \oplus \dots,$$

whose term A_q can be identified with the subspace of degree q homogeneous polynomial vector fields. We return to equation (1).

A simpler equation. Solving equation (1) means to figure out whether the given function α belongs to the image of the linear operator $w \mapsto Lw = wF$.

Our first step will be to solve equation (1) only for the leading terms of α. So if the expansion of α starts with the terms of order p,

$$\alpha = \alpha_p + \widetilde{\alpha}, \qquad \alpha_p \in M_p, \quad \widetilde{\alpha} \in \mathfrak{m}^{p+1},$$

we will try to solve equation (1) modulo terms of order $p+1$, that is, we look for a vector field $v = v(p)$ that provides the order p terms of α (modulo higher order terms) via the linear operator L,

$$\alpha_p = Lv \mod \mathfrak{m}^{p+1}. \tag{2}$$

This problem is simpler but equivalent to the initial one. Indeed, if we are able to find a vector field v that solves equation (2), then we can reduce the initial equation $\alpha = Lw$ to the equation

$$\alpha - Lv = L\widetilde{w}$$

on the unknown field $\widetilde{w} = w - v$ with the new function $\alpha - Lv$ on the left hand side whose expansion starts with the terms of order at least $p+1$. Proceeding by induction we compute all the terms of the required field w. Thus, our main goal is to resolve equation (2).

Homogeneous decomposition of F and L. To resolve equation (2), it is convenient to decompose the series F into homogeneous components

$$F = f_s + f_{s+1} + \ldots, \quad f_i \in M_i.$$

In this way, we decompose L into homogeneous operators that respect the graded structures, that is, $L = L_0 + L_1 + L_2 + \ldots$, with $L_r(A_{k-s}) \subset M_{k+r}$. Since the index of L_r is (traditionally) shifted by $-s$, for each $k \in \mathbb{Z}$ we shall write $\overline{k} := k - s$. Thus we have $L_r(A_{\overline{k}}) \subset M_{k+r}$:

$$\begin{tikzcd}
\ldots & \oplus & A_{\overline{k-1}} \arrow[dd, "L_0"] \arrow[ddrr, "L_1"] & \oplus & A_{\overline{k}} \arrow[dd, "L_0"] \arrow[ddrr, "L_1"] \arrow[ddrrrr, "L_2"] & \oplus & A_{\overline{k+1}} & \oplus & A_{\overline{k+2}} & \oplus & \ldots \\
 & & & & & & & & & \ldots & \\
\ldots & \oplus & M_{k-1} & \oplus & M_k & \oplus & M_{k+1} & \oplus & M_{k+2} & \oplus & \ldots
\end{tikzcd}$$

Each M_p contains the images of several homogeneous operators L_r and each $A_{\overline{k}}$ is the domain of several operators L_r. In the above diagram we have indicated with arrows only few of these operators.

Filtrations in the spaces $A_{\overline{k}}$ and M_p. We want to cover the biggest possible subspace of M_p applying L, but doing that we shall avoid to create terms of lower degree (in M_{p-1}, M_{p-2},...). To achieve it, when applying L to $A_{\overline{k}}$ we are going to restrict each component L_r of L to a subspace $A^r_{\overline{k}} \subset A_{\overline{k}}$ (defined below). In this way every space $A_{\overline{k}}$ will have a decreasing filtration of subspaces $(A_{\overline{k}} = A^0_{\overline{k}}) \supset A^1_{\overline{k}} \supset A^2_{\overline{k}} \ldots$, on which $L_0, L_1, L_2, \ldots$ will be restricted. Let us give a precise definition of these subspaces and fix a notation for the corresponding restricted operators and for their images.

Since L_0 has no preceding operator, for each $A_{\overline{k}}$ we note $A^0_{\overline{k}} := A_{\overline{k}}$ and

$$L^k_0 := L_0 : A^0_{\overline{k}} \longrightarrow M_k\,, \quad B^0_k := \operatorname{Im} L^k_0\,, \quad A^1_{\overline{k}} := \operatorname{Ker}\, L^k_0\,.$$

Next, we restrict the operator $L_1 : A_{\overline{k}} \to M_{k+1}$ to $A^1_{\overline{k}}$ (to avoid producing terms in M_k). We note this restriction by L^k_1 and define it modulo the image of the operator $L^{k+1}_0 : A^0_{\overline{k+1}} \longrightarrow M_{k+1}$ (see the above diagram), that is

$$L^k_1 : A^1_{\overline{k}} \longrightarrow M_{k+1}/B^0_{k+1}\,, \quad A^2_{\overline{k}} := \operatorname{Ker}\, L^k_1 \quad \text{and}$$

$$B_j^1 := (B_j^0 + \operatorname{Im} L_1^{j-1}) \subset M_j\,, \quad j \geqslant 0.$$

Now, we can define these objects inductively

$$L_r^k : A_{\overline{k}}^r \longrightarrow M_{k+r}/B_{k+r}^{r-1}\,, \qquad A_{\overline{k}}^{r+1} := \operatorname{Ker}\, L_r^k \quad \text{and}$$

$$B_j^r := (B_j^{r-1} + \operatorname{Im} L_r^{j-r}) \subset M_j\,, \quad j \geqslant r-1.$$

Hence besides the filtration of $A_{\overline{k}}$, each M_j is filtered by the increasing sequence of subspaces $B_j^0 \subset B_j^1 \subset B_j^2 \subset \ldots \subset M_j$. The main idea to define the operator L_r^k is that its source space must be the kernel of the "preceding" operator L_{r-1}^k, while its target space is defined modulo the accumulated images of all preceding operators that target in M_{k+r} $(L_{r-1}^{k+1}, L_{r-2}^{k+2}, \ldots, L_0^{k+r})$.

Finally, to solve equation (2) reduces to determine if $\alpha_p \in M_p$ can be covered by the (finite sum of) images of the operators that target in M_p

$$\begin{tikzcd}
\cdots & \oplus & A_{\overline{p}-3} \arrow[drrrrrr, "L_3^{p-3}"'] & \oplus & A_{\overline{p}-2}^2 \arrow[drrrr, "L_2^{p-2}"] & \oplus & A_{\overline{p}-1}^1 \arrow[drr, "L_1^{p-1}"] & \oplus & A_{\overline{p}} \arrow[d, "L_0^p"] & \oplus & \cdots \\
\cdots & \oplus & M_{p-3} & \oplus & M_{p-2} & \oplus & M_{p-1} & \oplus & M_p & \oplus & \cdots .
\end{tikzcd}$$

The algorithm. We check if $\alpha_p \in (B_p^0 = \operatorname{Im} L_0^p)$; if yes eq. (2) is solved. If not, we pass to the second step: we check if $\alpha_p \in (B_p^1 = \operatorname{Im} L_0^p + \operatorname{Im} L_1^{p-1})$; if yes eq. (2) is solved. If not, the 3rd step is to check if

$$\alpha_p \in (B_p^2 = \operatorname{Im} L_0^p + \operatorname{Im} L_1^{p-1} + \operatorname{Im} L_2^{p-2})$$

and so on. At each step one has to calculate the subspace $A_{\overline{p}-r}^r$ on which the operator L_r^{p-r} is defined.

Concretely, one first replaces the difficult linear ("differential") operator L in the right hand side of equation (2) by its first approximation $L_0^p : A_{\overline{p}} \to M_p$. If $\alpha_p \in \operatorname{Im} L_0^p$, that is, if there is a vector field $v \in A_{\overline{p}}$ such that

$$\alpha_p = L_0^p v\,,$$

then this vector field provides a solution of equation (2) because $L_0^p v$ and Lv differ by terms of degree greater than p.

But if this approximation is not sufficient, $\alpha \notin \operatorname{Im} L_0^p$, it might still happen that, replacing the operator L by its approximation $L_0^p + L_1^{p-1} + \cdots + L_r^{p-r}$ for some $r > 0$, equation (2) could be resolved with a vector of the form

$v = v_0 + v_1 + \cdots + v_r$ with $v_i \in A^i_{\overline{p}-i}$. In order to account such a possibility we proceed by induction in r. For $r = 1$ we look for a vector $v_1 \in A^1_{\overline{p}-1}$ (that is, $L_0^{p-1} v_1 = 0$) such that

$$\alpha_p - L_1^{p-1} v_1 \in \operatorname{Im}\left(L_0^p : A_{\overline{p}} \to M_p\right).$$

If such a vector field exists, then choosing v_0 satisfying $\alpha_p - L_1^{p-1} v_1 = L_0^p v_0$ we obtain

$$L(v_1 + v_0) = L_0^{p-1} v_1 + (L_0^p v_0 + L_1^{p-1} v_1) + \cdots = \alpha_p \mod \mathfrak{m}^{p+1},$$

which resolves equation (2). In other words, the second approximation is sufficient if α_p represents a vector in the image of the operator

$$L_1^{p-1} : A^1_{\overline{p}-1} \to M_p / B_p^0\,, \quad \text{where} \quad B_p^0 = \operatorname{Im}\left(L_0^p : A^0_{\overline{p}} \to M_p\right).$$

If α_p is not in the image of this operator, then the higher approximations are needed. This process is finite because $r \leqslant \overline{p} + 1$ and because, with the growth of r, the finite dimensional vector spaces $A^r_{\overline{q}}$ and B^r_p stabilise: $A^r_{\overline{q}} = A^{r+1}_{\overline{q}}$ and $B^r_p = B^{r+1}_q$.

One usually notes by A^∞_q and B^∞_p the limiting spaces, the symbol ∞ meaning a sufficiently large integer. Equation (2) is resolvable if and only if the function $\alpha_p \in M_p$ belongs to B^∞_p, that is, belongs to B^r_p for some r.

Remark. All these algebraic simple facts prove, in general, less than the necessary results for our problem in the algebra of power series. Namely, one starts from a filtered space $V^0 \supset V^1 \supset \ldots$ like $\mathbb{C}[[x]] \supset \mathfrak{m} \supset \mathfrak{m}^2 \supset \ldots$ in our example, but one "solve" the equation (calculate the homology,…) in a different space,

$$\widehat{V} = \bigoplus (V^k / V^{k+1}).$$

In calculus we can start from the holomorphic or smooth functions, obtaining the "solutions" in the class of formal series only: The convergence should be checked in the analytic case (it fails in some problems), the work in the smooth case being more complicated.

In the topological applications there is a similar difficulty: The formal algebraic answer of the spectral sequence method provides only the solution (homology,…) in the associated graded objects ($\widehat{V}$ above), rather than in the initial filtered ones. Fortunately, the results are equivalent in many cases, but one need to prove this equivalence.

A filtration (similar to that of our $\{\mathfrak{m}^p\}$ example) in the spaces of chains of the total space of a fibration (like the example of the unit tangent vectors of a surface) is provided by the following construction of J. Leray.

11.4 Leray's construction of spectral sequences

Consider a fibration $\pi : X^{m+n} \to B^m$ with typical fibre F^n . The following algorithm allows to compute the homology groups of the total space X^{m+n}, relating them to those of the base space and of the fibre. It is similar to the above algorithm to solve the homological equation.

First, one decomposes the base space B^m into simpler parts of different dimensions, say a cell partition (or a triangulation). This provides a filtration

$$(Y_{-1} = \varnothing) \subset Y_0 \subset Y_1 \subset Y_2 \subset \ldots \subset (Y_m = B^m), \tag{1}$$

where Y_0 is a finite subset, and Y_k is the union of Y_{k-1} with a finite number of disjoint cells of dimension k whose boundary simplices belong to Y_{k-1}. Thus Y_k is the k-skeleton of B. Next, filtration (1) induces a filtration

$$(X_{-1} = \varnothing) \subset X_0 \subset X_1 \subset \ldots \subset (X_m = X^{m+n})$$

of X by the subspaces $X_p = \pi^{-1}(Y_p)$. The boundary operator respects this filtration $\partial C_k(X_p) \subset C_{k-1}(X_p)$ (since Y_p is closed). Thus a cycle of X_p is a cycle of $X_{p+1} \subset X$. Hence the group $H_d(X)$ has an induced filtration

$$h_d(X_{-1}) = 0 \subset h_d(X_0) \subset h_d(X_1) \subset \ldots \subset h_d(X_m) = H_d(X)\,,$$

where $h_d(X_p)$ is the group formed by the homology classes in $H_d(X)$ that are obtained from the cycles in X_p (i.e., $h_d(X_p) = \operatorname{Im}(H_d(X_p) \to H_d(X))$ is the group of d-dimensional cycles in X_p modulo the group of boundaries in X).

It must be clear that the spectral sequence method does not provide this filtered homology group, but only its *associated graded group*:

$$\bigoplus_{p=0}^{m} h_d(X_p)\big/h_d(X_{p-1})\,,$$

which, in many cases, is isomorphic to the homology $H_d(X)$ itself (for example, if the coefficient group is a field). However, it does not always recover the torsion part of the homology (being necessary some extra arguments).

Graded chain groups. Defining the *degree* of a chain σ of X as the smallest p such that σ is a chain of X_p, the pth summand $h_d(X_p)\big/h_d(X_{p-1})$ can be seen as a subgroup of the "group of homogeneous chains of degree p". To describe $h_d(X_p)\big/h_d(X_{p-1})$ in terms of the chains of X one constructs a subdivision of X compatible with that of the base B^m in the following way: each d-dimensional cell of X is projected by π onto a cell of the subdivision of B^m, of any dimension not bigger than d. Moreover, the projection of the cells should be an "affine" fibration in some coordinate system.

Any such compatible subdivision provides a graduation in the group of d-dimensional chains of X :

$$C_d = C_{0,d} \oplus C_{1,d} \oplus \ldots \oplus C_{d,d}\,,$$

where $C_{p,d}$ is the *group of homogeneous d-dimensional chains of degree p* (generated by the d-cells whose projection is a p-dimensional cell of B^m, that is, the d-dimensional cells of $X_p \setminus X_{p-1}$). Thus, similarly to polynomials, any d-dimensional chain σ of degree p in X is a sum $\sigma = \sigma_p + \sigma_{p-1} + \ldots + \sigma_0$ (of homogeneous terms $\sigma_k \in C_{k,d}$) whose "leading term" σ_p is not zero.

Truncated homologies. Any two homology classes in $h_d(X_p)$ represented by two cycles with the same leading term (of degree p),

$$\sigma = \sigma_p + \sigma_{p-1} + \ldots + \sigma_0\,, \qquad \sigma' = \sigma_p + \sigma'_{p-1} + \ldots + \sigma'_0\,,$$

represent the same class in $h_d(X_p)\big/h_d(X_{p-1})$ because $\sigma - \sigma'$ is a cycle representing a class in $h_d(X_{p-1})$. That is, all d-cycles of degree p having the same leading term represent the same class in the quotient $h_d(X_p)\big/h_d(X_{p-1})$. Thus we can use the "truncated cycle" given by this common leading term to represent the corresponding class in $h_d(X_p)\big/h_d(X_{p-1})$. Furthermore, the leading term of any sum of d-cycles of degree p is equal to the sum of the leading terms of those cycles. Hence the group $h_d(X_p)\big/h_d(X_{p-1})$ is isomorphic to the group of truncated d-cycles of degree p modulo the truncated boundaries of degree p. Our goal is to calculate this group.

The above graduation in the chain groups automatically decomposes the boundary operator ∂ into homogeneous operators that respect the graduation, that is, $\partial = \partial_0 + \partial_1 + \partial_2 + \ldots$, where the rth differential, ∂_r, is a boundary operator in X^{m+n} that when applied to a cell belonging to X_p it takes into account the parts of its boundary belonging only to X_{p-r}, but neglecting those which live in X_{p-r-1}. Hence ∂_r lowers the degree by r and the dimension by 1, $\partial_r(C_{p,d}) \subset C_{p-r,d-1}$:

$$\begin{tikzcd}
\ldots \arrow[r, phantom, "\oplus"] & {} & C_{p-2,d} \arrow[r, phantom, "\oplus"] & {} & C_{p-1,d} \arrow[r, phantom, "\oplus"] & {} & C_{p,d} \arrow[dllll, "\partial_2"] \arrow[dll, "\partial_1"] \arrow[d, "\partial_0"] \arrow[r, phantom, "\oplus"] & {} & C_{p+1,d} \arrow[dll, "\partial_1"] \arrow[d, "\partial_0"] \arrow[r, phantom, "\oplus"] & {} & \cdots \arrow[r, phantom, "="] & {} & C_d \arrow[d, "\partial"] \\
\ldots \arrow[r, phantom, "\oplus"] & {} & C_{p-2,d-1} \arrow[r, phantom, "\oplus"] & {} & C_{p-1,d-1} \arrow[r, phantom, "\oplus"] & {} & C_{p,d-1} \arrow[r, phantom, "\oplus"] & {} & C_{p+1,d-1} \arrow[r, phantom, "\oplus"] & {} & \cdots \arrow[r, phantom, "="] & {} & C_{d-1}\,.
\end{tikzcd}$$

Hence, similarly to polynomials, the $(r+1)$th differential of any chain of degree r is zero.

Truncated boundaries. If a given cycle of degree p is a boundary, $\sigma = \partial u$, then the degree of u is necessarily bigger than $p-1$, unless $u=0$. Assuming that its degree is $p-1+r$ for some $r \geq 0$, $u = u_{p+r-1} + u_{p+r-2} + \dots$, the equation $\sigma = \partial u$ can be written as the following system of homogeneous equations (see the above diagram)

$$\begin{gathered} \partial_0 u_{p+r-1} = 0, \qquad \partial_1 u_{p+r-1} + \partial_0 u_{p+r-2} = 0\,, \quad \dots \quad , \\ \partial_{r-2} u_{p+r-1} + \dots + \partial_0 u_{p+1} = 0, \\ \partial_{r-1} u_{p+r-1} + \partial_{r-2} u_{p+r-2} + \dots + \partial_0 u_p = \sigma_p\,, \end{gathered} \tag{2}$$

where σ_p is the leading term of σ. Thus we define the subgroup $B^r_{p,d} \subset C_{p,d}$ to be the group of leading terms of boundaries ∂u such that the degree of u is bigger than $p-1$ at most by r. That is, $\sigma_p \in B^r_{p,d}$ if there exists a collection of homogeneous $(d+1)$-dimensional chains $u_{p+r-1}, u_{p+r-2}, \dots, u_p$ satisfying equations (2). In this way, we get an increasing sequence of subgroups

$$0 = B^0_{p,d} \subset B^1_{p,d} \subset B^2_{p,d} \subset \dots \subset C_{p,d}\,,$$

which stabilises, and the limiting group $B^{d-p+1}_{p,d} = B^{d-p+2}_{p,d} = \dots = B^\infty_{p,d}$ is exactly the group of d-dimensional truncated boundaries of degree p.

Truncated cycles. A chain of degree p, $\sigma = \sigma_p + \sigma_{p-1} + \dots + \sigma_0$, is a cycle if and only if

$$\partial_0 \sigma_p = 0, \;\; \partial_1 \sigma_p + \partial_0 \sigma_{p-1} = 0, \;\; \dots, \;\; \partial_p \sigma_p + \partial_{p-1}\sigma_{p-1} + \dots + \partial_0 \sigma_0 = 0. \tag{3}$$

Equations (3) impose conditions on the leading term σ_p. For each $r \geqslant 0$ we wish to take into account the first r of these conditions simultaneously. Namely, we say that $\sigma_p \in Z^r_{p,d}$, $r = 0,1,2,\dots$, if there exists a collection of homogeneous chains $\sigma_{p-1}, \dots, \sigma_{p-r+1}$ such that the first r equations of (3) are satisfied. Thus, (since $r=0$ means no condition) for $r = 0,1,2$ we have

$$Z^0_{p,d} = C_{p,d}, \qquad Z^1_{p,d} = \mathrm{Ker}\,(\partial_0), \qquad Z^2_{p,d} = \partial_1^{-1} \mathrm{Im}\,(\partial_0 : C_{p-1,d} \to C_{p-1,d-1})\,,$$

where ∂_1 is restricted to $\mathrm{Ker}\,(\partial_0) = Z^1_{p,d}$. We thus obtained a decreasing sequence of subgroups

$$C_{p,d} = Z^0_{p,d} \supset Z^1_{p,d} \supset Z^2_{p,d} \supset \dots$$

that stabilises. Its limiting group $Z^{p+1}_{p,d} = Z^{p+2}_{p,d} = \cdots = Z^{\infty}_{p,d}$ is exactly the group of truncated cycles. Summarising, we have

$$0 = B^0_{p,d} \subset B^1_{p,d} \subset B^2_{p,d} \subset \cdots \subset B^{\infty}_{p,d} \subset Z^{\infty}_{p,d} \subset \cdots \subset Z^2_{p,d} \subset Z^1_{p,d} \subset Z^0_{p,d} = C_{p,d}.$$

Main goal. Our desired group of leading terms of the d-dimensional cycles of degree p (modulo homological equivalence) is the quotient of the limiting groups $Z^{\infty}_{p,d}$ and $B^{\infty}_{p,d}$:

$$E^{\infty}_{p,d} := \frac{Z^{\infty}_{p,d}}{B^{\infty}_{p,d}} \quad \left(\approx \frac{h_d(X_p)}{h_d(X_{p-1})} \right),$$

and hence the graded group associated to $H_d(X)$ is isomorphic to $\bigoplus_{p=0}^{m} E^{\infty}_{p,d}$.

Our main goal is to compute $E^{\infty}_{p,d}$ by successive approximations.

Definition. The r-th approximation of the group $E^{\infty}_{p,d}$ is the quotient group

$$E^r_{p,d} := \frac{Z^r_{p,d}}{B^r_{p,d}}, \quad r = 0, 1, 2, \ldots \tag{4}$$

The 0-th term $E^0_{p,d}$ of this sequence of approximations is just the group of chains $C_{p,d}$. Observing these groups $E^r_{p,d}$ satisfy $d \geqslant p \geqslant 0$, we extend their definition to any $p, d \in \mathbb{Z}$ agreeing that $E^r_{p,d} = 0$ unless $d \geqslant p \geqslant 0$.

Higher differential homomorphism. To pass from the group $Z^r_{p,d}$ to $Z^{r+1}_{p,d}$ by induction, we need to consider the linear combination of boundaries

$$b = \partial_r \sigma_p + \partial_{r-1}\sigma_{p-1} + \cdots + \partial_0 \sigma_{p-r} \tag{5}$$

and to impose the condition that this combination is equal to zero. Naively, we might consider (5) as a 'higher differential homomorphism' from $Z^r_{p,d}$ to $C_{p-r,d-1}$. But remark that the chains $\sigma_{p-1}, \ldots, \sigma_{p-r+1}$ participating in the definition of the group $Z^r_{p,d}$ are not uniquely defined by the leading term σ_p, we just require the existence of *any* collection of chains satisfying the corresponding equations. It is easy to see that the indeterminacy of the element defined by (5) belongs to the group $B^r_{p-r,d-1}$. In other words, the operator ∂_r and expression (5) induce a well-defined homomorphism

$$Z^r_{p,d} \to C_{p-r,d-1}/B^r_{p-r,d-1},$$

whose kernel is the group $Z^{r+1}_{p,d}$. The essential fact on the approximating groups $E^r_{p,d}$ is that this homomorphism and the semi-exactness property of the initial boundary operator, $\partial^2 = 0$, determine an 'rth order' operator

$$\partial_r : E^r_{p,d} \to E^r_{p-r,d-1}$$

satisfying $\partial_r^2 = 0$. [Be warned: we use the same notation ∂_r.] Hence each group $E^r_{p,d}$ can be taken as a 'chain group' that enters in a chain complex

$$\cdots \to E^r_{p+r,d+1} \xrightarrow{\partial_r} E^r_{p,d} \xrightarrow{\partial_r} E^r_{p-r,d-1} \to \cdots .$$

The key property of the sequence of quotient groups (4) is the

Theorem. *The* $(r+1)$*th term of the sequence* (4) *of approximating quotient groups is naturally isomorphic to the homology of its* r*th term,*

$$E^{r+1}_{p,d} \approx \frac{\mathrm{Ker}\,(\partial_r : E^r_{p,d} \to E^r_{p-r,d-1})}{\mathrm{Im}\,(\partial_r : E^r_{p+r,d+1} \to E^r_{p,d})}. \tag{6}$$

Hence, in principle, *all groups* $E^r_{p,d}$ *can be computed recursively.*

To give a recursive algorithm, we construct a sequence E^r $(r \geqslant 0)$ of planar lattices of groups, called *pages*, as follows. For each $r \geqslant 0$ we assemble the groups $E^r_{*,*}$ into the lattice E^r, putting the group $E^r_{p,d}$ in the spot (p, d).

The *spectral sequence* has as his rth term the E^r page together with the boundary operator $\partial_r : E^r_{p,d} \to E^r_{p-r,d-1}$.

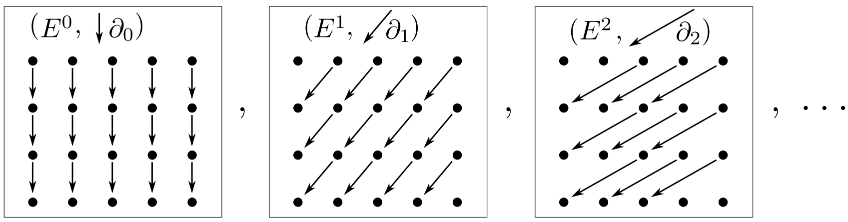


Figure 11.8: Spectral sequence pages together with their boundary operators.

Spectral sequence algorithm. It consists in recursively computing the spectral sequence pages using (6): Knowing the E^r page, (6) enables to compute the E^{r+1} page (the term $E^{r+1}_{p,d}$ being the homology defined by ∂_r at $E^r_{p,d}$). For example, in the 0th step, one starts with the E^0 page (consisting of the chain groups $E^0_{p,d} = C_{p,d}$) to get E^1. Etc.

Roughly speaking, to pass from the E^r page to E^{r+1}, the operator ∂_r acts on E^r as a sieve that gets rid of a part of the chains. Namely, one selects only the chains that form the kernel of ∂_r (the ∂_r-cycles), considering them modulo the ∂_r-boundaries. The process stabilises when the operators ∂_r are unable to filter nothing more.

We shall shift to the standard index notation, used everywhere, by setting $q = p - d$ and using (p, q) instead of (p, d) [unfortunately the index $q =$ dimension $-$ degree is a combination of two quantities of different nature]. In the new notation the differentials and homology groups become

$$\partial_r : E^r_{p,q} \to E^r_{p-r,q+r-1} \quad \text{and} \quad E^{r+1}_{p,q} = \frac{\operatorname{Ker}\left(\partial_r : E^r_{p,q} \to E^r_{p-r,q+r-1}\right)}{\operatorname{Im}\left(\partial_r : E^r_{p+r,q-r+1} \to E^r_{p,q}\right)} .$$

Hence the operator arrows of Fig. 11.8 change according to Fig. 11.9, the d-dimensional "homogeneous homology groups" lie along the diagonal $p+q = d$ and the graded group associated to $H_d(X)$ is isomorphic to $\bigoplus_{p+q=d} E^{\infty}_{p,q}$.

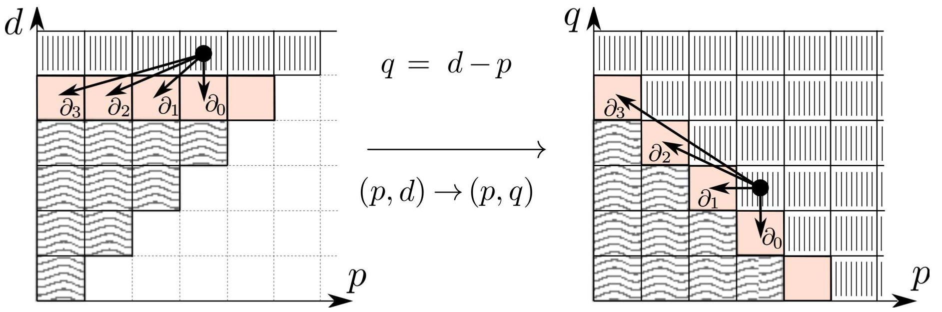


Figure 11.9: Geography of the spectral sequence differentials in the (p, d) (old) and (p, q) (new) notations. The differential ∂_r applies only to the page E^r.

The E^1 and E^2 pages have a useful explicit interpretation :

The E^1 **page**. Over each p-dimensional cell of B^m, the fibred space defines a "cylinder", which is a trivial fibration over that cell, being thus homotopy equivalent to the typical fibre F because cells are contractible.

By definition, the ∂_0-boundary of a chain σ of that cylinder is the part of $\partial\sigma$ contained in the cylinder (providing the usual homology in the cylinder). Thus a $(p+q)$-dimensional ∂_0-homology class over the chosen p-cell is represented by a homology class in $H_q(F)$, labeled by that p-cell. Therefore the E^1 page is formed by the groups of p-dimensional chains of the base space B^m with coefficients in the homology groups $H_q(F)$ of the fibre F^n :

$$E^1_{p,q} = C_p(B, H_q(F)) .$$

The E^2 page. The first differential, ∂_1, is a boundary operator in X^{m+n} that when applied to a chain over a chosen p-cell (i.e., belonging to X_p) it takes into account the parts of its boundary belonging only to X_p and to X_{p-1}, but neglecting those which live in X_{p-2}. If the fibration is *homologically trivial*[*], the homology group $E^2_{p,q}$ can be presented as the pth homology group of the base space B^m with coefficients in the qth homology group of the fibre F^n,

$$E^2_{p,q}(X,\mathbb{Z}) \approx \frac{\operatorname{Ker}(\partial_1 : E^1_{p,q} \to E^1_{p-1,q})}{\operatorname{Im}(\partial_1 : E^1_{p+1,q} \to E^1_{p,q})} = H_p(B, H_q(F,\mathbb{Z}))\,.$$

As we see, the E^2 page is uniquely determined by the homology of the base and of the fibre, being possible to start the algorithm at the E^2 page. For non homologically trivial fibrations, the description of E^2 requires an additional notion not considered in this book. In both cases, one can show that the subsequent pages and the higher differentials are independent of the cellular partition (triangulation) used to construct the spectral sequence.

Theorem. *For a trivial fibration* $(X = B \times F) \to B$ *the spectral sequence stabilises at* E^2 : $E^2_{p,q} = E^\infty_{p,q}$, *that is, all differentials* ∂_r *are zero for* $r \geqslant 2$. *Moreover, each homology group of the total space* X *is isomorphic (non-canonically) to its graded version provided by* E^2

$$H_d(X,\mathbb{Z}) \approx \bigoplus_{p+q=d} H_p(B, H_q(F,\mathbb{Z}))\,.$$

Example. Compute the homologies of $X = \mathbb{S}^1 \times M^2$, where M^2 is a genus $g > 1$ surface. We use that $H_0(M^2,\mathbb{Z}) = \mathbb{Z}$, $H_1(M^2,\mathbb{Z}) = \mathbb{Z}^{2g}$, $H_2(M^2,\mathbb{Z}) = \mathbb{Z}$, $H_{k\geqslant 3}(M^2,\mathbb{Z}) = 0$; $H_0(\mathbb{S}^1,G) = G$, $H_1(\mathbb{S}^1,G) = G$, $H_{k\geqslant 2}(\mathbb{S}^1,G) = 0$, and take the fibration $\mathbb{S}^1 \times M^2 \to \mathbb{S}^1$:
$H_0(X) \approx H_0(\mathbb{S}^1,\mathbb{Z}) = \mathbb{Z}$,
$H_1(X) \approx H_0(\mathbb{S}^1,\mathbb{Z}^{2g}) \oplus H_1(\mathbb{S}^1,\mathbb{Z}) = \mathbb{Z}^{2g} \oplus \mathbb{Z}$,
$H_2(X) \approx H_0(\mathbb{S}^1,\mathbb{Z}) \oplus H_1(\mathbb{S}^1,\mathbb{Z}^{2g}) \oplus H_2(\mathbb{S}^1,\mathbb{Z}) = \mathbb{Z} \oplus \mathbb{Z}^{2g} \oplus 0 = \mathbb{Z} \oplus \mathbb{Z}^{2g}$ and
$H_3(X) \approx H_0(\mathbb{S}^1,0) \oplus H_1(\mathbb{S}^1,\mathbb{Z}) \oplus H_2(\mathbb{S}^1,\mathbb{Z}^{2g}) \oplus H_3(\mathbb{S}^1,\mathbb{Z}) = 0 \oplus \mathbb{Z} \oplus 0 \oplus 0 = \mathbb{Z}$.

Exercise. In p. 419, we have stated that the Betti numbers of the n-dimensional torus, $\mathbb{T}^n = (\mathbb{S}^1)^n$, are the binomial coefficients, $b_k(\mathbb{T}^n) = C^k_n$ (namely $H_k(\mathbb{T}^n) \approx \mathbb{Z}^{C^n_k}$). Prove it by induction using that $\mathbb{T}^n = \mathbb{T}^{n-1} \times \mathbb{S}^1$.

---

[*] This means that the action of the group $\pi_1(B)$ on the homology groups of F (induced by the lifting property) is trivial.

Exercise. Compute the homology groups of the direct product manifold $X = \mathbb{R}\mathrm{P}^2 \times \mathbb{R}\mathrm{P}^2$. Hint: Use the homologies calculated on p. 419.

Answer. $H_0(X) = \mathbb{Z},\ H_1(X) = \mathbb{Z}_2 \oplus \mathbb{Z}_2,\ H_2(X) = \mathbb{Z}_2,\ H_3(X) = \mathbb{Z}_2,\ H_{k\geqslant 4}(X) = 0.$

We have to indicate that there exist non-trivial fibrations whose spectral sequence stabilises at the 2nd page. However, for most fibrations the subsequent pages $E^{r\geqslant 3}$ provide homological information on their non-triviality. These pages are obtained from the differentials $\partial_{r\geqslant 2}$ whose computation depends on the properties of each particular fibration.

An example of that is the operator D computed on pp. 430-430 (Fig. 11.10).

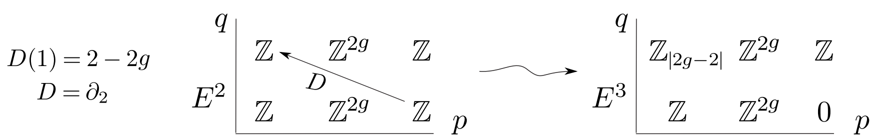


Figure 11.10: The spectral sequence for the homology groups of the 3-manifold of the unit tangent vectors of a surface of genus $g \neq 1$.

In this case the spectral sequence provides the graded version

$$E^3_{1,0} \oplus E^3_{0,1} = \mathbb{Z}^{2g} \oplus \mathbb{Z}_{2g-2},$$

proving that the group $H_1(X, \mathbb{Z})$, in the exact sequence

$$0 \to \mathbb{Z}_{2g-2} \to H_1(X, \mathbb{Z}) \to \mathbb{Z}^{2g} \to 0,$$

is isomorphic to $\mathbb{Z}^{2g} \oplus \mathbb{Z}_{2g-2}$ for $g \neq 1$.

Problem. Let X^{m+n} be the total space of the fibration $\pi : X^{m+n} \to B^m$ with spherical fibres $\mathbb{S}^n$. Prove the exactness of the natural sequence of groups and homomorphisms:

$$\ldots \to H_{p+1}(B) \xrightarrow{D_{p+1}} H_{p-n}(B) \xrightarrow{\alpha_p} H_p(X) \xrightarrow{\pi_*} H_p(B) \xrightarrow{D_p} H_{p-n-1}(B) \to \ldots \quad (7)$$

which provides the graded version $H_{p-1}(B)/(\operatorname{Im} D_{p+1}) \oplus \operatorname{Ker} D_p$ of $H_p(X)$:

$$0 \to \frac{H_{p-n}(B)}{\operatorname{Im} D_{p+1}} \to H_p(X) \to \operatorname{Ker} D_p \to 0.$$

Solution. In (7) the homomorphism α_p sends a $(p-n)$-cycle in the base space to its full preimage under π, which is a p-cycle in the fibration space X; the homomorphisms $D_\bullet$ are given by the spectral sequence differential ∂_{n+1} (like D in Fig. 11.10). The only relevant differential here is ∂_{n+1} because $E^2_{p,q} = 0$ for $q \neq 0, n$ (implied by $H_q(\mathbb{S}^n) = 0$ for $q \neq 0, n$). The E^{n+1} page, which leads to (7), is shown in Fig. 11.11.

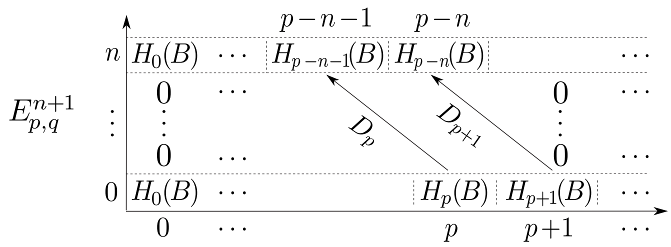


Figure 11.11: The spectral sequence of the fibration $\pi : X^{m+n} \to B^m$ with spherical fibres $\mathbb{S}^n$. Here $E^2_{p,0} = E^2_{p,n} = H_p(B)$ follows from $H_0(\mathbb{S}^n) = H_n(\mathbb{S}^n) = \mathbb{Z}$.

Problem. Find a similar exact sequence for the fibrations $\pi : X^{m+n} \to B^m$ with base $B^m = \mathbb{S}^m$, supposing the homologies of the typical fibre F^n are known.

Hint. Taking into account that $E^2_{p,q} = 0$ for $p \neq 0, m$ (because $H_p(\mathbb{S}^m) = 0$ for $p \neq 0, m$) the only non trivial differential is determined by the vertical lines $p = 0$ and $p = m$.

11.4 bis Spectral sequence of a filtered space

Observing that the chain graduation of p. 437 is artificial because it depends on the chosen compatible subdivision, here we present spectral sequences in a more invariant and shorter way. In order to allow the reader to compare directly this presentation with the above constructions, in this section, we will use our previous (non standard) indices (p, d).

Here, no fibration structure is required for the space X (of course, all what follows can be applied if X has a fibration structure). We only assume to have an arbitrary filtration in X (with no a priori relation to the cell partition of X) :

$$(X_{-1} = \varnothing) \subset X_0 \subset X_1 \subset \ldots \subset (X_m = X).$$

Such filtration induces a filtration in the group of d-dimensional chains :

$$\varnothing = F_{-1}C_d \subset F_0 C_d \subset F_1 C_d \subset \cdots \subset F_m C_d = C_d \,,$$

where the subgroup $F_p C_d$ is generated by the d-dimensional cells contained in X_p. For any subgroup $A \subset C_d$, let us denote by $\mathrm{Gr}_p A$ its 'pth graded quotient',

$$\mathrm{Gr}_p A = \frac{A \cap F_p C_d}{A \cap F_{p-1} C_d} = \frac{A \cap F_p C_d + F_{p-1} C_d}{F_{p-1} C_d} \subset \frac{F_p C_d}{F_{p-1} C_d} = \mathrm{Gr}_p C_d.$$

Applying this to the subgroups of cycles $Z_d \subset C_d$ and boundaries $B_d \subset Z_d$ we obtain the natural isomorphisms

$$h_d(X_p)/h_d(X_{p-1}) = \mathrm{Gr}_p Z_d/\mathrm{Gr}_p B_d.$$

For every cycle $c \in Z_d \cap F_p C_d$ representing an element $[c]$ of $\mathrm{Gr}_p Z_d$ the condition $\partial c = 0$ can be expressed by saying that ∂c has negative filtration. Thus for an integer $r \geqslant 0$ we define the group $Z^r_{p,d} \supset \mathrm{Gr}_p Z_d$ by the weaker condition that the filtration of ∂c is smaller than that of c by at least r, that is, $\partial c \in F_{p-r} C_{d-1}$.

Similarly, any boundary $b \in B_d \cap F_p C_d$ representing an element $[b]$ of $\mathrm{Gr}_p B_d$ satisfy $b \in \mathrm{Im}\,(\partial : C_{d+1} \to F_p C_d)$. So we define the subgroup $B^r_{p,d} \subset \mathrm{Gr}_p B_d$ by the condition that b is the boundary of some chain whose filtration is bigger than that of b by at most $r-1$, that is, $b \in \mathrm{Im}\,(\partial : F_{p+r-1} C_{d+1} \to F_p C_d)$. The two sequences of subgroups we have just defined, $Z^r_{p,d}$ $(r = 0, 1, 2, \ldots)$ and $B^r_{p,d}$ $(r = 0, 1, 2, \ldots)$, satisfy

$$0 = B^0_{p,d} \subset B^1_{p,d} \subset B^2_{p,d} \subset \cdots \subset \mathrm{Gr}_p B_d \subset \mathrm{Gr}_p Z_d \subset \cdots \subset Z^2_{p,d} \subset Z^1_{p,d} \subset Z^0_{p,d} = \mathrm{Gr}_p C_d,$$

$$Z^r_{p,d} = \mathrm{Gr}_p(\partial^{-1} F_{p-r} C_{d-1}), \qquad B^r_{p,d} = \mathrm{Gr}_p(\partial(F_{p+r-1} C_{d+1})).$$

Both sequences stabilise: $Z^r_{p,d} = \mathrm{Gr}_p Z_d$ for $r > p$ and $B^r_{p,d} = \mathrm{Gr}_p B_d$ for $r > d-p$ (stabilisation usually occurs earlier). So one usually denotes

$$Z^\infty_{p,d} := \mathrm{Gr}_p Z_d, \quad B^\infty_{p,d} := \mathrm{Gr}_p B_d.$$

Therefore the sequence of quotient groups $E^r_{p,d} := Z^r_{p,d}/B^r_{p,d}$ $(r = 0, 1, 2, \ldots)$ stabilises at the desired limiting group

$$E^\infty_{p,d} := \frac{Z^\infty_{p,d}}{B^\infty_{p,d}} = \frac{\mathrm{Gr}_p Z_d}{\mathrm{Gr}_p B_d} = \frac{h_d(X_p)}{h_d(X_{p-1})}.$$

The existence of 'higher differential operators' $\partial_r : E^r_{p,d} \to E^r_{p-r,d-1}$ satisfying $\partial_r^2 = 0$ provides several chain complexes $\cdots \to E^r_{p+r,d+1} \xrightarrow{\partial_r} E^r_{p,d} \xrightarrow{\partial_r} E^r_{p-r,d-1} \to \cdots$. The key property of the sequence of quotient groups $E^r_{p,d}$ $(r = 0, 1, 2, \ldots)$ is that its $(r+1)$th term is naturally isomorphic to the homology of its rth term,

$$E^{r+1}_{p,d} \approx \frac{\mathrm{Ker}\,(\partial_r : E^r_{p,d} \to E^r_{p-r,d-1})}{\mathrm{Im}\,(\partial_r : E^r_{p+r,d+1} \to E^r_{p,d})}. \tag{1}$$

The differential ∂_r is induced by the differential ∂ of the original chain complex of the space X. If c is a chain representing an element of the group $E^r_{p,d}$, then its boundary ∂c is not necessarily zero, but, by definition, it has filtration at most $p-r$, so, we can take it as a representative of the element $\partial_r c$ in the group $E^r_{p-r,d-1}$. The verification of the correctness of this definition and of the validity of (1) is an exercise of linear algebra.

The construction of the E^r pages and of the spectral sequence algorithm (using (1)) is identical to that of the previous section. However, here, the E^2 page has no special interpretation. The standard notation uses the above (p, q) indices (Fig. 11.9).

Although the spectral sequence method is mostly used to the explicit computation of homology groups, it is also used to prove theoretical theorems. For example, it provides one of the simplest proofs that singular and cellular homologies are equivalent for cell spaces. Spectral sequences are also used to prove that two different topological spaces have isomorphic homology groups, without explicit computation of those homology groups. We have only shown the beginning of the theory. Spectral sequence theory have been developed in several directions in topology and homological algebra.

11.5 Morse theory

Morse theory of generic smooth functions provides one of the most geometric and important methods to investigate the homologies of smooth manifolds.

Let $f : M \to \mathbb{R}$ be a generic smooth function on a compact n-dimensional manifold M. The main idea is to describe the modifications of the sets $M_c = \{z \in M : f(z) \leqslant c\}$ when the parameter c is growing (Fig. 11.12).

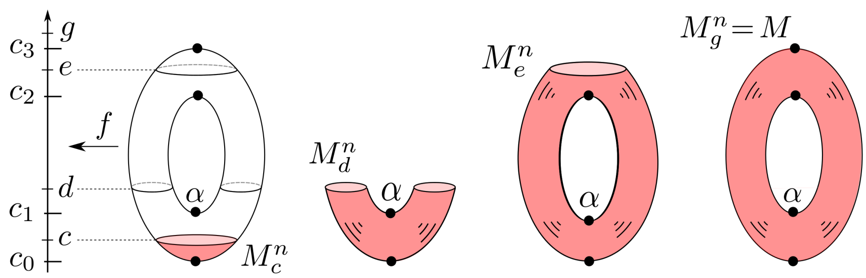


Figure 11.12: Structural modifications of the sets M_c at the critical points.

When $c < \min f$ the set M_c is void. At $c = c_0 = \min f$ the set M_c consists of the point where c is reached by f. For $c = c_0 + \varepsilon$ with $\varepsilon > 0$ the set M_c is diffeomorphic to a ball B^n of $\mathbb{R}^n$ (by the Morse lemma, p. 538).

The situation between two consecutive critical values $c_k < c < c_{k+1}$ is stable: The manifolds (with boundary) M_c are diffeomorphic to each other for all values c in this open interval. The diffeomorphisms are provided by the gradient flow of the function f, normalised to reach $L_v f = \text{const}$.

The main point is to understand that the manifold (with boundary) $M_{c_k+\varepsilon}$ is obtained from $M_{c_k-\varepsilon}$ by a "Morse modification of handle attaching" (explained below) at the critical point c_k – Fig. 11.13.

The difference between $M_{c_k-\varepsilon}$ and $M_{c_k+\varepsilon}$ is essential only in a small neighbourhood of the critical point z_k whose critical value is c_k * – Fig. 11.14.

*Outside this neighbourhood one may identify $M_{c_k+\varepsilon}$ to $M_{c_k-\varepsilon}$ by a diffeomorphism provided by the gradient flow.

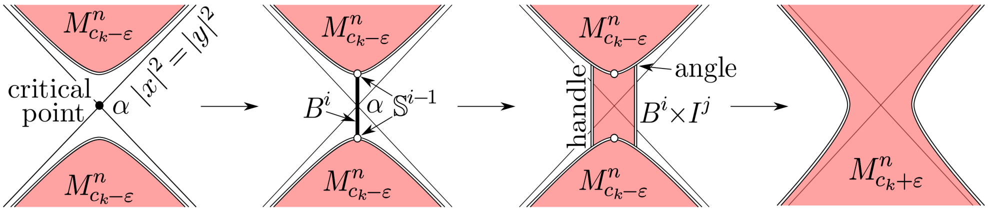


Figure 11.13: Handle attaching at a critical point of index $i = 1$.

In this neighbourhood we can use a coordinate system in which f has the normal form of Morse theorem,

$$f = c_k + x_1^2 + \ldots + x_j^2 - y_1^2 - \ldots - y_i^2\,,$$

where $i + j = n$. The number i of negative squares is called the *index* of the critical point z_k. The index i has the following influence in the modification that we want to study: There is a homotopy equivalence

$$M_{c_k+\varepsilon} \sim M_{c_k-\varepsilon} \cup B^i, \tag{μ}$$

where the i-dimensional ball B^i is glued to the 'lower' manifold $M_{c_k-\varepsilon}$ by an embedding

$$\varphi : (\mathbb{S}^{i-1} = \partial B^i) \to \partial M_{c_k-\varepsilon}.$$

The modification (μ) is called the *handle attaching*.

To replace the homotopy equivalence (μ) by a diffeomorphism, one has to replace the i-dimensional ball B^i by its product with a complementary fibre I^{n-i}, embedding the boundary fibres $\mathbb{S}^{i-1} \times I^{n-i}$ to $M_{c_k-\varepsilon}$ along the $j = n - i$ x-directions of the Morse normal form (3rd picture of Fig. 11.13) smoothening then the resulting angles (4th picture of Fig. 11.13).

In Fig. 11.13 the critical point z has index $i = 1$ and the ball B^1 is a segment which is attached to $M_{c_k-\varepsilon}$, connecting two points of $\partial M_{c_k-\varepsilon}$ (along which $f = c_k - \varepsilon$) in a small neighbourhood of z. In Fig. 11.14 $i = 2$.

The formal proof of formula (μ) can be read in the nice book of J. Milnor [102], but we hope the reader would be able to understand it from Fig. 11.12 and can prove it by himself.

The main point now is to understand the sphere $\mathbb{S}^{i-1}$ $(x = 0,\ |\,y\,|^2 = \varepsilon)$ on the boundary of the 'lower' manifold $M_{c_k-\varepsilon}$, to which the handle B^i is attached in (μ).

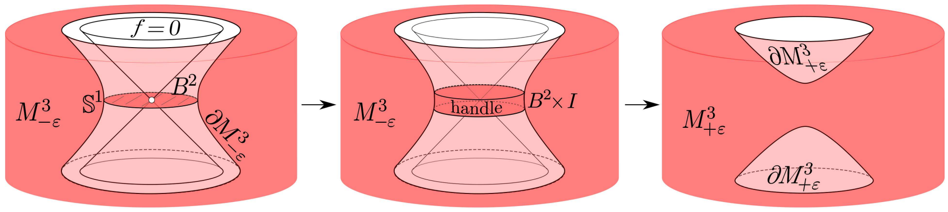


Figure 11.14: The Morse modification of index $i = 2$ for $f = x^2 - y_1^2 - y_2^2$.

The homology of the manifold (with boundary) M_c changes at each attachment of a handle in the following way: A new generating chain B^i of dimension i is attached at the modification of index i, and its boundary ∂B^i is some linear combination of the generating $(i-1)$-chains on the 'lower' manifold $M_{c_k-\varepsilon}$ (defined by the inclusion φ of the sphere $\mathbb{S}^{i-1} = \partial B^i$, oriented as the boundary of the oriented disc B^i which defines the new generator of dimension i).

We must repeat this homology version of the handle attaching, as many times as the number of critical points of the chosen generic smooth function $f : M \to \mathbb{R}$, whose critical points are non degenerate and whose critical values are all different. At the end, we get the *Morse complex* of the function f:

$$\ldots \to C_i \stackrel{\partial_i}{\to} C_{i-1} \to \ldots \ ,$$

where the spaces of Morse chains are $C_i = \mathbb{Z}^{m_i}$, and m_i is the number of critical points of index i of the chosen function.

The homologies of this complex are the homologies of the manifold M, independently of the special choice of the generic function f.

This fact is proved by the exact sequences of the sequence of attachments of handles. Say, for the function of Fig. 11.12, the groups of Morse chains are

$$C_0 = \mathbb{Z}\,\alpha, \quad C_1 = \mathbb{Z}\,z + \mathbb{Z}\,w, \quad C_2 = \mathbb{Z}\,\omega,$$

and the boundary homomorphisms are

$$\partial\alpha = 0, \quad \partial z = \alpha - \alpha = 0, \quad \partial w = \alpha - \alpha = 0, \quad \partial\omega = z + w - z - w = 0.$$

To compute the boundary in the general situation, it suffices to use the gradient* dynamics of the function f in the following way:

*To define the gradient vector field, we need to fix some Riemannian metric on M.

The phase curves of the field $\operatorname{grad} f$ that start along the y-directions of the Morse normal form at the critical point z of index i form an i-dimensional "descending" manifold of the critical point z (we may choose the y-directions to be orthogonal to the x-direction by reducing the Morse normal form function to the orthogonal normal form).

The similar "ascending gradient lines" (starting along the x-directions) of a critical point w of index $i-1$ form a smooth manifold of dimension $j+1 = n-i+1$ in M.

Choosing a non-critical level c, $f(w) \leqslant c \leqslant f(z)$, we find in the smooth $(n-1)$-dimensional manifold $N = f^{-1}(c)$ (of codimension 1 in M) the representative A of the descending manifold of the critical point z, which has dimension $i-1$, and the representative B of the ascending $(n-i+1)$-manifold of the critical point w, which has dimension $n-i$.

The submanifolds A^{i-1} and B^{n-i} intersect, in N^{n-1}, at a finite number of points if f is generic. The coefficient of the Morse chain w in the Morse chain ∂z is equal to the number of these intersection points $A^{i-1} \cap B^{n-i}$, counting them according to the orientations of the chosen generators z and w.

These connecting lines which descend from one critical point to an other are called *instantons*. The coefficient of w in ∂z is equal to the sum of the contributions of all instantons from z to w.

Remark. The invariance under homeomorphisms of the Morse complex homologies has not been proved above, and the proof is not so easy.

PROBLEM. Compute the Morse complexes of the 28 Milnor spheres (p. 110).

Hint. There exist functions with only two (non degenerate) critical points on each of these smooth non-diffeomorphic exotic "spheres". This fact implies that all the 28 Milnor spheres are homeomorphic to the standard one.

Poincaré Duality. Let M be a compact connected oriented smooth n-dimensional manifold. Comparing the generic function f with the generic function $-f$, one observes that the critical points of index i for f are the critical points of index $n-i$ for $-f$. The descending gradient line of f which connects z to w coincides (up to the orientation reversal) with the descending gradient line of the function $-f$, connecting w to z.

We conclude that the Morse complexes of the functions f and $-f$ are conjugate to one another (in the sense of linear algebra), the group of i-chains for the function f is dual to the group of $(n-i)$-chains for the function $-f$.

Poincaré Duality Theorem. *For any compact oriented closed n-dimensional manifold M, the Betti numbers $b_i(M)$ and $b_{n-i}(M)$ are equal.*

This provides another proof of the

Corollary. *The Euler characteristic of any odd dimensional manifold is zero.*

Indeed, since the Euler characteristic is the alternated sum of the Betti numbers, the Poincaré duality $b_i(M) = b_{n-i}(M)$ implies that $\chi(M^n) = 0$ for odd n because each term appears twice in the sum, but with opposite signs.

Poincaré duality theorem is very close to a slightly more general fact, also discovered by Poincaré, that we shall discuss now.

11.6 Intersection duality and cohomology

Let us associate to each pair a, b of cycles of complementary dimensions, $a \in H_i,\ b \in H_{n-i},$ their intersection number (a, b). This defines the bilinear *intersection pairing map*

$$H_i(M) \times H_{n-i}(M) \to \mathbb{Z}\,.$$

Thus to each element $b \in H_{n-i}$ the intersection pairing associates the linear map $\beta : H_i \to \mathbb{Z},$ whose value at $a \in H_i$ is the number $\beta(a) = (a, b),$ which depends only on the homology classes of the cycles.

Intersection duality theorem. *The intersection pairing is non degenerate: Every linear map $H_i \to \mathbb{Z}$ coincides with the map β for a suitable element b of H_{n-i}.*

A convenient reformulation of this result uses the notion of cohomology.

Cohomology theories For each homology theory there is a corresponding "cohomology theory" constructed as follows. One starts considering the groups of chains $C_i = C_i(M, \mathbb{Z})$ used in the definition of homology. Then the Abelian group $C^i(M, \mathbb{Z})$ formed by the linear maps $C_i \to \mathbb{Z}$ is called the *group of i-dimensional cochains*. These groups form the *cochain complex*

$$\ldots \to C^{i-1} \overset{\delta_i}{\to} C^i \overset{\delta_{i+1}}{\to} C^{i+1} \to \ldots\ ,$$

where the *coboundary* operator $\delta_i : C^{i-1} \to C^i$ is conjugate (in the sense of linear algebra) to the boundary operator $\partial_i : C_i \to C_{i-1}$ (the matrices are transposed to one another).

Similarly to the cycles, boundaries and homology groups, one defines the subgroups of *cocycles* $Z^k = \operatorname{Ker} \delta_{i+i}$ and *coboundaries* $B^k = \operatorname{Im} \delta_{k-1}$ in C^k leading to the quotient *cohomology groups*

$$H^k(M, \mathbb{Z}) = Z^k / B^k.$$

This definition makes sense due to the property $\delta_{i+1} \circ \delta_i = 0$.

Replacing the integer-valued cochains with cochains taking values in some Abelian group one arrives to the similar definition of cohomology with coefficients in this group.

If the group of coefficients is a field, then the k-th cohomology group is the vector space dual to the vector space of k-th homology.

In the case of integer coefficients the free parts of $H_k(M, \mathbb{Z})$ and $H^k(M, \mathbb{Z})$ are dual to one another, in particular, these groups have the same ranks b_k. The relation between the torsion subgroups in the integer homology and cohomology groups is more complicated. According to the so called *universal coefficient formula*, knowing all integer homology or cohomology groups is sufficient to recover all homology and cohomology with any given coefficient group, but we shell not discuss these algebraic questions here.

The intersection duality theorem provides a natural isomorphism

$$H^i(M, \mathbb{Z}) \approx H_{n-i}(M, \mathbb{Z})$$

that holds taking the torsion into account for every smooth closed oriented n-manifold M.

These Poincaré results remain true for topological manifolds because the natural homomorphisms discussed above are invariant under homeomorphisms.

We shall not prove these topological theorems in the non-smooth case. They are naturally generalised to the construction of the *intersection ring* whose multiplication operation* is the *intersection of the cycles*,

$$H_i(M, \mathbb{Z}) \times H_j(M, \mathbb{Z}) \to H_{i+j-n}(M, \mathbb{Z}).$$

Example. The intersection ring of $\mathbb{C}\mathrm{P}^n$ is the ring of truncated polynomials $\mathbb{Z}[t]/t^{n+1}$, where $t \in [\mathbb{C}\mathrm{P}^{n-1}]$. In this ring t^{n+1} is 0.

*We write it for the $\mathbb{Z}$-coefficients, but other rings ($\mathbb{Q}$, $\mathbb{Z}_p$, $\mathbb{R}$, $\mathbb{C}$) can be also used.

The dual operation in the cohomologies is the bilinear map

$$H^i(M,\mathbb{Z}) \times H^j(M,\mathbb{Z}) \to H^{i+j}(M,\mathbb{Z}),$$

called *cohomological multiplication.*

This bilinear operation provides the structure of an associative and super-commutative algebra $H^*\times H^* \to H^*$ to the graded space of the cohomologies of M of all dimensions $H^*(M) = \bigoplus H^i(M)$. The *super-commutativity* property means $ab = (-1)^{k\ell}ba$ if $a \in H^k(M)$, $b \in H^\ell(M)$.

Example. The cohomological algebra of $\mathbb{C}\mathrm{P}^n$ has one multiplicative generator of dimension 2: $[\mathbb{C}\mathrm{P}^{n-1}] \in H^2(\mathbb{C}\mathrm{P}^n)$.

The defining relation is $[\mathbb{C}\mathrm{P}^{n-1}]^{n+1} = 0$. The cohomology class $[\mathbb{C}\mathrm{P}^{n-1}]$ is defined as the linear function on $H_2(\mathbb{C}\mathrm{P}^n)$, whose value on any 2-cycle is the intersection number of that 2-cycle with the hyperplane $\mathbb{C}\mathrm{P}^{n-1}$.

Although this *cohomological algebra* was originated from the Poincaré intersection ring of an oriented manifold, it is more general – being well defined (and invariant under homeomorphisms) for complexes more general than smooth or topological manifolds. For example, the intersection rings are missing for a singular algebraic variety with any singularities or for any singular polyhedron, but the cohomological algebras are well defined.

11.7 De Rham cohomology

There is an analytical version of cohomology theory, which relates it to the smooth differential forms. It is very natural to think on this version if one takes into account the duality of the differential forms to the chains, which we have discussed in Ch. 6, p. 213-215.

Definition. Let M be an n-dimensional smooth manifold. The *i-th de Rham cohomology group of M* is defined as

$$H^i_{DR}(M,\mathbb{R}) := \frac{(\text{closed } i\text{-forms on } M)}{(\text{exact } i\text{-forms on } M)}. \qquad (\lambda)$$

Denote the vector space of the smooth i-forms on M by $\Omega_i(M,\mathbb{R})$ and the exterior derivative of the forms by $d : \Omega_i \to \Omega_{i+1}$. We can thus write the definition (λ) in the form

$$H^i_{DR}(M,\mathbb{R}) = \frac{\mathrm{Ker}\,(d : \Omega_i(M,\mathbb{R}) \to \Omega_{i+1}(M,\mathbb{R}))}{\mathrm{Im}\,(d : \Omega_{i-1}(M,\mathbb{R}) \to \Omega_i(M,\mathbb{R}))}.$$

Example. For the torus $M = \mathbb{T}^2 = \{\alpha \in \mathbb{R} \pmod{2\pi}, \beta \in \mathbb{R} \pmod{2\pi}\}$ we have computed the exterior derivatives (Ch. 6, p. 200):

$$d(Pd\alpha + Qd\beta) = \left(\frac{\partial Q}{\partial \alpha} - \frac{\partial P}{\partial \beta}\right)(d\alpha \wedge d\beta), \qquad d(f(\alpha,\beta)) = \frac{\partial f}{\partial \alpha}d\alpha + \frac{\partial f}{\partial \beta}d\beta.$$

The numerator in (λ) for the covering plane $\mathbb{R}^2$, instead of $\mathbb{T}^2$, is equal to the denominator by the Poincaré lemma (Ch. 6, p. 219). Therefore, on the torus the quotient space (λ) represents the differentials of the "multivalued functions" on $\mathbb{T}^2$ (these differentials being periodic on $\mathbb{R}^2$), modulo the differentials of the periodic functions. So, we may choose as the generators of $H^1_{DR}(\mathbb{T}^2, \mathbb{R})$ the classes of the forms $d\alpha$ and $d\beta$. This shows that

$$H^1_{DR}(\mathbb{T}^2, \mathbb{R}) = \mathbb{R}^2.$$

Problem. Calculate the de Rham cohomology group $H^k_{DR}(\mathbb{T}^n, \mathbb{R})$ of the n-torus $\mathbb{T}^n = (\mathbb{S}^1)^n$.

Answer. $H^k_{DR}(\mathbb{T}^n, \mathbb{R}) = \mathbb{R}^{b_k}$, where $b_k = C^k_n$. Compare to the calculation of the Betti numbers of the torus $\mathbb{T}^n$ on page 419.

Problem. Calculate the de Rham cohomology group $H^2_{DR}(\mathbb{R}^3 \setminus 0, \mathbb{R})$ of the complement to the point $0 \in \mathbb{R}^3$.

Answer. $H^2_{DR}(\mathbb{R}^3 \setminus 0, \mathbb{R}) = \mathbb{R}$. The generator ω is the flux form of the vector field of the gravitational attraction (see Ch. 6, p. 191 for the expression in coordinates of the form ω).

Problem. Calculate the de Rham cohomology group $H^1_{DR}(\mathbb{R}^2 \setminus 0, \mathbb{R})$.

Answer. $H^1_{DR}(\mathbb{R}^2 \setminus 0, \mathbb{R}) = \mathbb{R}$. The generating 1-form is (see Ch. 6, p. 218)

$$\omega = \frac{xdy - ydx}{x^2 + y^2}.$$

11.8 De Rham theorem and Čech cohomology

Let M be a manifold without boundary, not necessarily compact.

De Rham theorem. $H^*_{\mathrm{DR}}(M) \approx H^*(M, \mathbb{R})$.

In the proof given below, we use the *Čech cohomology groups* constructed in the following way.

Fix a locally finite cover of M by open subsets $M = \bigcup U_\alpha$ such that the intersection $U_{\alpha_0} \cap \cdots \cap U_{\alpha_p}$ of any finite collection of subsets of the cover is either empty or *contractible* (diffeomorphic to an open ball in Euclidean space). Along this section we fix such locally finite cover of M.

Definition. A *Čech p-cochain* is a map that, to any non-empty intersection $U_{\alpha_0} \cap \cdots \cap U_{\alpha_p} \neq \varnothing$ of an ordered collection of $p+1$ pairwise different subsets $U_{\alpha_0}, \ldots, U_{\alpha_p}$ of the cover, associates a real number $f_{\alpha_0,\ldots,\alpha_p} \in \mathbb{R}$. The order of the indices (i.e., of the subsets) is important. We assume the cochains are *anti-symmetric* with respect to permutations of indexes $\alpha_0, \ldots, \alpha_p$.

In other words, a Čech p-cochain is a map that, to each non-empty intersection of $p+1$ ordered distinct subsets of the cover, it associates a constant function on that intersection.

Example. A Čech 0-cochain is a map that associates to each subset of the cover a constant function on it.

The p-cochains form a vector space denoted by $C^p = C^p(M, \mathbb{R})$.

The *Čech cohomology* of M is defined as the cohomology of the cochain complex

$$0 \longrightarrow C^0 \xrightarrow{\delta} C^1 \xrightarrow{\delta} C^2 \xrightarrow{\delta} \ldots,$$

where the coboundary operator δ is defined by

$$(\delta f)_{\alpha_0,\ldots,\alpha_{p+1}} = \sum_{j=0}^{p+1} (-1)^j f_{\alpha_0,\ldots,\alpha_{j-1},\alpha_{j+1}\ldots,\alpha_{p+1}}.$$

Problem. Consider the complement to a point in the plane, $M = \mathbb{R}^2 \setminus \{0\}$, and take a cover of it by three open half-planes, such that each two of these half-planes intersect while all of them together have empty intersection. Write the Čech complex of this cover of $M = \mathbb{R}^2 \setminus \{0\}$, and calculate the corresponding Čech homologies.

Solution. Since the cover consists of three subsets, any Čech 0-cochain is a triple of constants (f_0, f_1, f_2), and hence the group of Čech 0-cochains is $C^0 = \mathbb{R}^3$. Again, since, up to order, the number of different non-empty intersections of the subsets of the cover is three and that the Čech cochains are anti-symmetric, any Čech 1-cochain is a triple of constants, and the group of Čech 1-cochains is $C^1 = \mathbb{R}^3$. The corresponding Čech complex has the form

$$0 \longrightarrow (C^0 = \mathbb{R}^3) \xrightarrow{\delta} (C^1 = \mathbb{R}^3) \longrightarrow 0 \longrightarrow 0 \cdots .$$

The non-trivial coboundary operator has rank 2: its kernel is the "diagonal" line, which consists of the maps that associate to each subset of the cover the same constant ($f_0 = f_1 = f_2$). Therefore, the cohomologies of degree 0 and 1 of this complex are one-dimensional and they are trivial in all other degrees. This is what we expect because $\mathbb{R}^2 \smallsetminus \{0\}$ is homotopy equivalent to the circle $\mathbb{S}^1$.

Theorem 2. *The Čech cohomology of M is isomorphic to the simplicial cohomology of M.*

An example of the cover of $\mathbb{R}^2 \smallsetminus \{0\}$ by 5 half-plane is missing, but I have doubts if it worth to include it.

Proof. Consider a triangulation (i.e., simplicial partition) of M. To any vertex (a zero dimensional simplex) α we associate its *star* U_α: It is the union of all open simplices adjacent to α. The subsets U_α form an open cover of M. The intersection $U_{\alpha_0} \cap \cdots \cap U_{\alpha_p}$ is non-empty if and only if the points $\alpha_0, \dots, \alpha_p$ are vertices of some simplex of the partition. Thus, in this case the group of the Čech p-cochains is equal to the group of the corresponding simplicial p-cochains. Moreover, the Čech differential δ is equal to the usual coboundary operator on the simplicial complex. Therefore, both complexes coincide and they have isomorphic cohomologies. □

Remark. Different covers of a manifold could provide different Čech complexes for it, but, due to the isomorphism of the theorem, the Čech cohomology is independent of the choice of the cover (if the cover satisfies the contractibility property for non-empty intersections of the covering subsets).

The last theorem implies that to prove the de Rham theorem it is sufficient to establish an isomorphism between the de Rham and Čech cohomologies.

Example. Let us show the computations leading to the isomorphism

$$H^1_{DR}(M, \mathbb{R}) \approx H^1(M, \mathbb{R}) \approx \mathbb{R}$$

for the manifold $M = \mathbb{R}^2 \smallsetminus \{0\}$ of the problem on p. 454.

A de Rham cohomology class is represented by a closed 1-form ω on M. We cannot guarantee the exactness of this form. However, by Poincaré lemma, this form is exact on each of the three half-planes U_0, U_1, and U_2 covering M. Therefore, there exist functions f_0, f_1, and f_2 such that f_α is defined in U_α and $df_\alpha = \omega|_{U_\alpha}$ for $\alpha = 0, 1, 2$.

Denote by $g_{\alpha,\beta} = f_\beta - f_\alpha$ the difference of these functions on the intersection of the half-planes $U_\alpha \cap U_\beta$. Then $dg_{\alpha,\beta} = df_\beta - df_\alpha = 0$. Therefore, the function $g_{\alpha,\beta}$ is constant; it can be treated as a Čech 1-cochain. This cochain is obviously a cocycle, it provides the necessary cohomology class.

Conversely, assume that we are given a Čech cohomology class represented by a cocycle $\{g_{\alpha,\beta}\}$. Then one can find functions f_0, f_1, and f_2 defined on the half planes U_0, U_1, and U_2, respectively, such that for any pair of indexes α and β the functions f_a and f_β differ on the intersection $U_\alpha \cap U_\beta$ by the constant $g_{\alpha,\beta}$, $f_\beta = f_\alpha + g_{\alpha,\beta}$ (see the lemma below). Although the functions f_α may differ on the intersection of the half-planes, their differentials $df_\beta = df_\alpha + dg_{\alpha,\beta} = df_\alpha$ coincide and determine a globally defined 1-form on M representing the necessary de Rham cohomology class.

The constructed maps

$$H^1_{DR}(M, \mathbb{R}) \to H^1(M, \mathbb{R}) \quad \text{and} \quad H^1(M, \mathbb{R}) \to H^1_{DR}(M, \mathbb{R})$$

are inverse to one another providing the necessary isomorphism.

Essentially, the proof of the de Rham theorem in the general case differs from the above arguments just by a more complicated notations. It uses the following generalisation of the notion of a Čech cochain.

Definition. A *Čech* (p, q)*-cochain* is a map that, to each non-empty intersection $U_{\alpha_0} \cap \cdots \cap U_{\alpha_p}$ of $p+1$ ordered distinct subsets of the cover, associates a smooth differential q-form on that intersection. We assume the cochain is anti-symmetric with respect to the ordering of the indices $\alpha_0, \ldots, \alpha_p$.

For p and q fixed, the set of (p, q)-cochains form a vector space denoted by $C^{p,q} = C^p(\Omega^q(M))$.

The spaces $C^{p,q}$ are related by the de Rham differential $C^{p,q} \xrightarrow{d} C^{p,q+1}$, which is just the exterior derivation $(d\omega)_{\alpha_0,\ldots,\alpha_p} = d\omega_{\alpha_0,\ldots,\alpha_p}$, and by the natural Čech differential $C^{p,q} \xrightarrow{\delta} C^{p+1,q}$, which is given by the formula above, but replacing the constant functions by the differential q-forms. These differentials evidently commute, $\delta d = d\delta : C^{p,q} \to C^{p+1,q+1}$, and satisfy $\delta^2 = 0$, $d^2 = 0$.

Lemma. *The sequence*

$$C^{p,0} \xrightarrow{d} C^{p,1} \xrightarrow{d} C^{p,2} \xrightarrow{d} \dots$$

is exact for any p. The sequence

$$C^{0,q} \xrightarrow{\delta} C^{1,q} \xrightarrow{\delta} C^{2,q} \xrightarrow{\delta} \dots$$

is exact for any q.

The first assertion follows from the Poincaré lemma, since we have supposed the contractibility of the intersections $U_{\alpha_0} \cap \dots \cap U_{\alpha_p}$ (this is the place where this condition is used!)

The second assertion is verified directly. For example, in the case $p = 1$, for any collection of differential forms $\{\omega_{\alpha,\beta}\}$ satisfying $\omega_{\alpha,\beta} + \omega_{\beta,\gamma} + \omega_{\gamma,\alpha} = 0$ we need to construct a collection of forms $\{\eta_\alpha\}$ such that $\omega_{\alpha,b} = \eta_\beta - \eta_\alpha$. This can be done (for all p) as follows. For a given cover we can find a *partition of unity*: a collection of C^∞ smooth functions ϱ_α on M such that ϱ_α is supported on U_α and $\sum_\alpha \varrho_\alpha \equiv 1$ (this is the place where we use the condition on the cover to be locally finite). Then for any δ-closed cochain $\omega \in C^{p,q}$ we define $\eta \in C^{p-1,q}$ by setting

$$\eta_{\alpha_0,\dots,\alpha_{p-1}} = \sum_\beta \varrho_\beta \, \omega_{\beta,\alpha_0,\dots,\alpha_{p-1}},$$

where the form $\varrho_\beta \, \omega_{\beta,\alpha_0,\dots,\alpha_{p-1}}$ extends to $U_{\alpha_0} \cap \dots \cap U_{\alpha_{p-1}}$ by zero; one verifies that $\omega = \delta\eta$. In the case $p = 1$, explicitly:

$$(\delta\eta)_{\alpha,\beta} = \eta_\beta - \eta_\alpha = \sum_\gamma \varrho_\gamma \, \omega_{\gamma,\beta} - \sum_\gamma \varrho_\gamma \, \omega_{\gamma,\alpha} = \sum_\gamma \varrho_\gamma \, \omega_{\alpha,\beta} = \omega_{\alpha,\beta}.$$

Now we construct the desired isomorphism $H^k_{\mathrm{DR}}(M, \mathbb{R}) \to H^k(M, \mathbb{R})$:

Proof of the de Rham Theorem. Remark that the group $\Omega^k(M)$ of k-cochains in the de Rham complex can be identified with the subgroup of $C^{0,k}$ consisting of δ-closed elements (indeed, any collection $\{\omega_\alpha\}$ of differential forms defined on U_α and satisfying $\omega_\alpha = \omega_\beta$ on $U_\alpha \cap U_\beta$ determines uniquely a globally defined form ω).

Similarly, the group $C^k(M)$ of k-cochains in the Čech complex can be identified with the subgroup $C^{k,0}$ consisting of d-closed elements (indeed, a

d-closed function on a contractible set is a constant). Consider the following diagram

$$
\begin{tikzcd}
\Omega^k(M) \arrow[r, hook] & C^{0,k} & & C^{1,k-1} & \cdots & C^{k-1,1} & & C^{k,0} & C^k(M,\mathbb{R}) \arrow[l, hook'] \\
& & C^{0,k-1} \arrow[ul, "d"] \arrow[ur, "\delta"'] & & & & C^{k-1,0} \arrow[ul, "d"] \arrow[ur, "\delta"'] & &
\end{tikzcd}
$$

Let a cohomology class in $H^k_{\mathrm{DR}}(M)$ be represented by a closed differential form ω. Denote by $\omega^{0,k}_\alpha$ the restriction of ω to U_α. The collection of all these forms $\omega^{0,k}_\alpha$ determines a Čech $(0,k)$-cochain, that is, an element of the group $C^{0,k}$. By Poincaré lemma, each form $\omega^{0,k}_\alpha$ is exact, $\omega^{0,k}_\alpha = d\eta_\alpha$, where the $(k-1)$-form η_α is defined on U_α only. Applying δ we obtain a cochain $\omega^{1,k-1} \in C^{1,k-1}$ given by $\omega^{1,k-1}_{\alpha,\beta} = \eta_\beta - \eta_\alpha$. Remark that the cochain $\omega^{1,k-1}$ is d-closed:

$$d\omega^{1,k-1}_{\alpha,\beta} = d\eta_\beta - d\eta_\alpha = (\omega - \omega)|_{U_\alpha \cap U_\beta} = 0.$$

Continuing in this way we define inductively the cochains

$$\omega^{i+1,k-i-1} = \delta d^{-1}\omega^{i,k-i} \in C^{i+1,k-i-1}, \quad i = 0, 1, 2, \dots,$$

which are all both d-closed and δ-closed. After k steps we obtain the cochain $\omega^{k,0} \in C^{k,0}$ representing the desired Čech cohomology class. The existence of the primitive $d^{-1}\omega^{i,k-i}$ at each step is guaranteed by the Poincaré lemma; the arbitrariness in the choice of the primitive leads to a cohomologous representative of the Čech cohomology class. The inverse map $H^k(M,\mathbb{R}) \to H^k_{\mathrm{DR}}(M)$ is defined similarly, by exchanging d and δ in the construction above. This completes the proof. □

Remark. The proof given above uses essentially nothing but the Poincaré lemma. One can give a similar proof to show the equivalence of other cohomology theories. For example, to show the isomorphism of the Čech and singular cohomology, we take as the group $C^{p,q}$ the group whose elements are given by collections of singular q-cochains $\omega_{\alpha_0,\dots,\alpha_p}$ on the space $U_{\alpha_0} \cap \cdots \cap U_{\alpha_p}$ given for all tuples of $p+1$ pairwise different subsets of the cover with non-empty intersection. Then the analogue of the lemma holds true and the whole proof of the isomorphism repeats word-by-word the arguments above.

Remark (Hypercohomology). Using the spectral sequence method, the presented proof can be reformulated as follows.

Consider the groups $\mathbf{C}^k = \bigoplus_{p+q=k} C^{p,q}$ and define $D = d + (-1)^p\delta : \mathbf{C}^k \to \mathbf{C}^{k+1}$. Then $D^2 = 0$, and we obtain a cochain complex

$$0 \to \mathbf{C}^0 \xrightarrow{D} \mathbf{C}^1 \xrightarrow{D} \mathbf{C}^2 \xrightarrow{D} \dots.$$

The cohomology $\mathbf{H}^*(M)$ of this complex is called the *hypercohomology* of M.

With the spectral sequence method, one computes the hypercohomology in several steps. In the first approximation we replace D by d ignoring the δ-summand of the differential; then we compute the correction coming from the δ-term in D, etc.

It follows from the lemma above that the first term $E_1^{p,q}$ is reduced just to one line: It is trivial for $q > 0$ and the line $E_1^{0,0} \to E_1^{1,0} \to E_1^{2,0} \to \dots$ is the complex for the computation of the Čech cohomology of M. Therefore, the second term of the spectral sequence, E_2, is the Čech cohomology and the higher differentials are trivial by dimensional reasons. Thus, the spectral sequence provides the isomorphism of the hypercohomology with the Čech cohomology of M.

Similarly, with exchanging d and δ, we construct another spectral sequence that provides the isomorphism of the hypercohomology with the de Rham cohomology. Finally, we obtain the required isomorphism

$$H^*_{\mathrm{DR}}(M, \mathbb{R}) \approx \mathbf{H}^*(M) \approx H^*(M, \mathbb{R}).$$

11.9 Swallowtails and complexity theory

The first homology groups $H_i(G_k)$ of the complement to the complex swallowtails

$$G_k = \{\lambda \in \mathbb{C}^k : x^{k+1} + \lambda_1 x^{k-1} + \dots + \lambda_k \ \text{ has no multiple root}\}$$

are presented in the following table

$k \diagdown i$	0	1	2	3	4	5	6	7	8
$1, 2$	$\lvert\mathbb{Z}\rvert$	$\lvert\mathbb{Z}\rvert$	0	0	0	0	0	0	0
$3, 4$	$\mathbb{Z}$	$\mathbb{Z}$	$\lvert\mathbb{Z}_2\rvert$	0	0	0	0	0	0
$5, 6$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}_2$	$\lvert\mathbb{Z}_2\rvert$	$\mathbb{Z}_3$	0	0	0	0
$7, 8$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\lvert\mathbb{Z}_6\rvert$	$\mathbb{Z}_3$	$\mathbb{Z}_2$	0	0
$9, 10$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_6$	$\lvert\mathbb{Z}_6\rvert$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_5$

The natural stabilisation $H_i(G_{k+1}) \approx H_i(G_k)$ occurs for $k \geqslant 2i-1$. The first stable groups in the table are written between two vertical bars: $|H|$.

G. Segal and J. May discovered that the stable cohomology groups $H^*(G_\infty)$ are isomorphic to that of the second loop space of $\mathbb{S}^3$: $H^*(G_\infty) \approx H^*(\Omega^2\mathbb{S}^3)$.

The natural isomorphisms $H_\bullet(G_{2n}) = H_\bullet(G_{2n-1})$ hold for any n. All the groups $H_i(G_k)$ are finite for $i \geqslant 2$.

The proofs of these results and of their generalisations (including the description of the cohomology rings) are published in: [6, 7, 8, 74, 114].

These results were used lately by S. Smale and V. Vassiliev for the estimation of the complexity of any algorithm which calculates the roots of the complex polynomials of degree n. The final Vassiliev estimations of these Smale complexities (which we shall not define here and which had been introduced earlier by A. Schwartz [113], who called them "fibration genus") are:

$$n - C\ln n \leqslant \text{complexity} \leqslant n.$$

The Smale-Vassiliev theory is discussed in [119, 130, 51, 52]. Chapter 2 of the book [134] contains a survey on the applications to algorithms and complexity theory.

Configurations of hyperplanes

Consider a finite collection of hyperplanes $\{z : \alpha_k(z) = 0\}$ in the complex vector space $\mathbb{C}^n$, and consider the complement of the configuration they form

$$G = \{z \in \mathbb{C}^n : \prod \alpha_k(z) \neq 0\}.$$

Its cohomology ring $H^*(G, \mathbb{Z})$ has the following description: It is generated by the 1-dimensional classes

$$\omega_k = \frac{1}{2\pi}\frac{d\alpha_k}{\alpha_k},$$

and a polynomial in these generators is cohomologous to 0 if and only if it is equal to 0.

Example (Diagonal hyperplanes). For the configuration formed by the C_n^2 diagonal hyperplanes defined by the equations $z_p = z_q$ in the coordinate system $(z_1, \ldots, z_n)$ in $\mathbb{C}^n$, the Poincaré series* of the cohomology ring of the complement to the union of all diagonal hyperplanes (graded by the dimensions of the cohomology classes) is equal to

$$p(t) = (1+t)(1+2t)\ldots(1+(n-1)t).$$

The value $p(1) = n!$ is equal to the number of components (called "Weyl chambers") into which the real diagonal hyperplanes subdivide the real space $\mathbb{R}^n$. The fundamental group of G, $\pi_1(G)$, is the coloured braid group of n strings (see Ch.5, p.169) while its higher homotopy groups are trivial.

Similar results hold for many other configurations of complex hyperplanes, which are obtained as the complexification of some configuration of real hyperplanes of $\mathbb{R}^n$. For instance, the Arnold studies of the previous example (for the case of the coloured braid group) were extended by Brieskorn to other generalised braid groups associated to any finite reflection group, and by Deligne to the complexification of configurations of hyperplanes that subdivide the real space into simplicial cones (which always happens for the real reflection groups), proving the vanishing of the higher homotopy groups of the complement to such complexified hyperplane configurations.

In all cases of complexified real configurations, the coefficients of the Poincaré polynomials count the number of connected components of the complement to the union of real hyperplanes in the real space (reflecting also the combinatorics of their intersections).

These theories are described, with many generalisations, in the book [106], which contains also an extensive bibliography.

The strange relation discussed above between the topology of the real domains defined by a combination of real hyperplanes and that of the complex domain, obtained from the complexified version of such combination, has the following general meaning.

The Betti numbers of a real algebraic manifold $M^m_{\mathbb{R}} \subset \mathbb{R}\mathrm{P}^n$ are related to the Betti numbers of the set of complex points $M^m_{\mathbb{C}} \subset \mathbb{C}\mathrm{P}^n$ (defined by

---

*By definition, the Poincaré series of G is the series

$$p(t) = \sum_{k=0}^{\infty} b_k t^k, \quad \text{where} \quad b_k = \dim H^k(G, \mathbb{R}).$$

the same algebraic equations) by the Smith inequality

$$\sum_{k=0}^{m} b_k(M_{\mathbb{R}}, \mathbb{Z}_2) \leqslant \sum_{k=0}^{2m} b_k(M_{\mathbb{C}}, \mathbb{Z}_2). \qquad (\mu)$$

In the case of the union of real hyperplanes (or of their complementary spaces) the Smith inequality (μ) becomes an equality.

Definition. A real algebraic variety is called *M-variety* (M from "maximal"), if the Smith inequality (μ) for it is an equality.

Example. The configurations of hyperplanes are M varieties.

Example. A smooth real algebraic curve of genus g in $\mathbb{R}\mathrm{P}^2$ is an M-curve if is has $g+1$ connected components ("ovals").

The ellipses are M-curves, the real elliptic curves are M-curves when they consist of two components (see Fig. 5.22), but those consisting of one component are not M-curves.

Hilbert announced that there exist exactly two non diffeomorphic M-curves of degree 6, but their number is three (see Fig. 5.18, on page 150).

The possible values of the Betti numbers of the M-manifolds are unknown, but the right hand side numbers do not depend on the particular choice of the non degenerate manifold, by the "Italian principle" of Ch. 5, p. 145.

For instance, for the hypersurfaces defined in $\mathbb{C}\mathrm{P}^n$ by one equation of fixed degree d, it suffices to compute the numbers $b_k(M_{\mathbb{C}})$ for one example, and it suffices to replace the highly degenerate equation $\prod_{k=1}^{d} \alpha_k(z) = 0$ of the union of d generic hyperplanes in $\mathbb{C}\mathrm{P}^n$ by the slightly perturbed equation.

Problem. Calculate the Betti numbers of a smooth complex hypersurface (of complex dimension $n-1$) in $\mathbb{C}\mathrm{P}^n$.

Answer. See Chapter 12 and the next section.

11.10 Characteristic classes (a review)

A *characteristic class* is a cohomology class associated in a natural way to some geometric structure on a manifold (vector bundle, smooth structure, embedding to some ambient space etc).

The main example that we will consider here of such a geometric structure is that of a vector bundle E over a manifold M.

A vector bundle is said to be *oriented* if the fibres are oriented in a continuous way with respect to the points of M.

Example. The so-called first Stiefel-Whitney class is an example of a characteristic class of a vector bundle E over M. This class is an element $w_1(E) \in H^1(M, \mathbb{Z}_2)$, whose non triviality is an obstruction to the orientability of the bundle E: it takes the value 1 or 0 (mod 2) on a closed loop, depending on whether the fibre of the bundle reverses or preserves its orientation along that loop.

Example. The Euler class $e(E) \in H^n(M)$ is an obstruction to the existence of a global nowhere vanishing section of an oriented rank n vector bundle E over a smooth manifold M. *It is Poincaré dual to the zero locus of a generic section of that bundle.* This means that one considers a generic section of the bundle E and takes the set Z formed by the zeros of that section. The fact that the section is generic implies that Z is a smooth submanifold of codimension n in M – here we identify M with the zero section of the bundle.

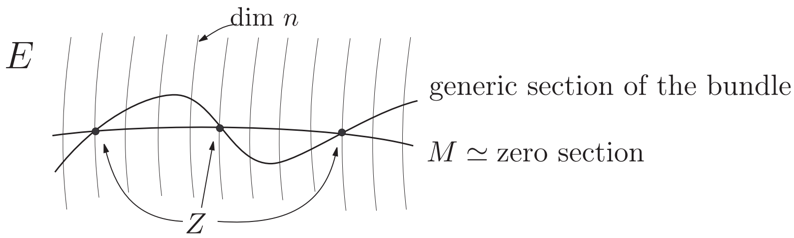


Figure 11.15: The Poincaré dual to the zero locus of a generic section.

One takes the Poincaré dual of Z, which is a degree n cohomology class: $[Z] \in H^n(M)$. Observe that there are two cases:

If the bundle is oriented, then Z is equipped with an induced natural co-orientation. In this case the Euler class is an integer cohomology class.

If one considers an arbitrary bundle, orientable or non orientable, then there is no (natural) co-orientation on Z. In this case the Euler class belongs to $H^n(M, \mathbb{Z}_2)$.

The above words 'in a natural way' have a explicit mathematical meaning that has to be made precise depending on the situation. The geometric structure under consideration can often be induced from a canonical one over a universal 'classifying space' of such structures. In such case, the ring of characteristic classes is just the cohomology ring of that classifying space.

The simplest closed theory of characteristic classes is that of rank n complex vector bundles – a complex vector bundle is a bundle whose fibre is equipped with a complex structure, but such a bundle can be defined over any real or complex manifold. The classifying space in this case is the complex Grassmannian $G^{\mathbb{C}}_{n,N}$, $N \gg 0$. The ring of characteristic classes (that is, the cohomology ring of the Grassmannian) is freely generated by certain generators $c_1, \dots, c_n$, $c_i \in H^{2i}(G^{\mathbb{C}}_{n,N})$, called *Chern classes*.

The actual choice of these generators c_i is fixed by the following properties:

First, by definition, the top Chern class $c_n(E)$ is the Euler class of that bundle. In particular, for a line bundle (that is, of rank one) $c_1(L) = e(L)$.

Next, one defines the class c_k as the only class such that in the particular case where E is the direct sum of n line bundles, $E = L_1 \oplus \dots \oplus L_n$, the class $c_k(E)$ is the kth elementary symmetric function in the classes $c_1(L_1), \dots, c_1(L_n)$. Namely, denoting these classes by $t_1 = c_1(L_1), \dots, t_n = c_1(L_n)$, this definition can be written in the following convenient way:

$$c(E) = (1 + t_1) \dots (1 + t_n),$$

where $c(E) = 1 + c_1(E) + \dots + c_n(E)$. This non-homogeneous class is often referred to as the *total Chern class*. Having in mind this formula, one extends sometimes the definition of Chern classes by setting $c_0(E) = 1$ and $c_k(E) = 0$ for $k < 0$ or $k > n = \mathrm{rk}(E)$. By the *splitting principle*, the above definition of Chern classes extends in a unique way to arbitrary complex vector bundles – that is, not necessarily admitting a split into a direct sum of n line bundles.

In general, the splitting principle claims that in all computations related to the Chern classes of arbitrary complex vector bundles, it is sufficient to make the computations only for the case when the bundles admit splitting as direct sums of line bundles. Namely, the classes appearing in the computations are symmetric with respect to $t_1, \dots, t_n$, and the expression for the resulting classes in terms of the elementary symmetric functions c_i is universal, i.e. it can be applied to the bundles that do not admit splitting into line bundles. So, using the splitting principle, one can easily compute the Chern classes of a direct sum of two vector bundles, tensor product, etc. For example, one gets immediately the *Whitney formula*

$$c(E \oplus F) = c(E)\, c(F),$$

i.e. $c_k(E \oplus F) = \sum_{i+j=k} c_i(E)c_j(F)$.

The flag bundle. The splitting principle is justified by the following considerations. First, note that a map $\varrho : N \to M$ to the base M of the bundle, induces a bundle $E' \to N$ with base N. This way of obtaining new bundles from a given bundle, is called "change of base operation". A condition that a class must satisfy to be a characteristic one is to have good behaviour under the change of base operation: The classes obtained in the induced bundle $E' \to N$ from the characteristic classes of the initial bundle $E \to M$ by mean of the homomorphism ϱ^*, must coincide with the characteristic classes of the induced bundle.

Next, one considers the manifold $F\ell(E)$ whose points are all possible complete flags taken on each fibre of $E \to M$, and takes the natural map $\varrho : F\ell(E) \to M$, whose fibre over a point $x \in M$ is the space of all the complete flags on the fibre V_x of the bundle $E \to M$. Using the map ϱ, one makes a change of base operation to get the induced vector bundle

$$\varrho^*(E) \to F\ell(E).$$

This bundle admits a splitting obtained from the complete flags (one uses the orthogonal complements of the flags to construct the splitting on lines).

Finally, one proves that the homomorphism $\varrho^* : H^*(M) \to H^*(F\ell(E))$ is injective. Then, all properties of a characteristic class of the initial vector bundle are equal to those of its image on $H^*(F\ell(E))$ because the injectivity of the homomorphism.

Hence, after a proper change of base, one can reduce the study of the characteristic classes of an arbitrary complex vector bundle to the case where the bundle is split.

Problem. Find the first Chern class of the tautological line bundle τ over the projective space $\mathbb{C}\mathrm{P}^m$.

Solution. Consider an arbitrary nonzero linear function on $\mathbb{C}^{m+1}$. Restricting this function to different lines one obtains a section of the dual bundle τ^*. The zero locus of this section is the projective hyperplane $\mathbb{C}\mathrm{P}^{m-1}$. Therefore, $c_1(\tau^*) = h$ where $h \in H^2(\mathbb{C}\mathrm{P}^m)$ is the cohomology class Poincaré dual to a hyperplane. Respectively, $c_1(\tau) = -h$.

Problem. Compute the Chern classes of the tangent bundle to the complex projective space $\mathbb{C}\mathrm{P}^m$.

Solution. There is a natural isomorphism

$$T\mathbb{C}\mathrm{P}^m \approx (\tau^*)^{m+1}/\mathbb{C}.$$

Indeed, an infinitesimal movement of a line $\ell \subset \mathbb{C}^{m+1}$ can be given by a linear operator $\ell \to \mathbb{C}^{m+1}$, that is, by a vector of the space $(\ell^*)^{m+1}$. In this space, the homotheties form a trivial subspace that preserves the line. Whence the isomorphism.

By Whitney formula we get from the isomorphism the equality

$$c(T\mathbb{C}\mathrm{P}^m) = \frac{(c(\tau^*))^{m+1}}{c(\mathbb{C})} = (1+h)^{m+1},$$

that is, $c_k(T\mathbb{C}\mathrm{P}^m) = C_{m+1}^k h^k$, where h is the class of a hyperplane.

Problem. Let M be a degree n generic algebraic hypersurface in $\mathbb{C}\mathrm{P}^m$ (given by a degree n homogeneous polynomial in $m+1$ variables with generic coefficients). Compute the Chern classes of the tangent bundle of M.

Solution. The tangent bundle TM is a sub-bundle of the restriction to M of the tangent bundle to the ambient projective space $\mathbb{C}\mathrm{P}^m$. The quotient bundle has rank one and is isomorphic to

$$T_M\mathbb{C}\mathrm{P}^m/TM \approx (\tau^*)^{\otimes n},$$

the nth tensor power of the bundle τ^* dual to the tautological one. It follows from the fact that the polynomial determining M can be treated as a section of the bundle $(\tau^*)^{\otimes n}$, and M can be identified with the zero locus of this section.

By Whitney formula we get from the above isomorphism the equality

$$c(TM) = \frac{(1+h)^{m+1}}{1+nh},$$

where $h \in H^2(M)$ is the cohomology class of a hyperplane section.

Problem. Find the Euler characteristic of the hypersurface M in the previous problem.

Solution. By definition of the top Chern class, the Euler characteristic is equal to the value of $c_{m-1}(TM)$ computed in the previous problem on the fundamental class $[M] = n[\mathbb{C}\mathrm{P}^{m-1}]$ of the manifold M:

$$\chi(M) = n\sum_{i=2}^{m+1} C_{m+1}^i(-n)^{i-2} = \frac{(1-n)^{m+1} - 1 - n(m+1)}{n}.$$

A more geometric approach leading to the same formula will be discussed in Chapter 12.

The classifying space of real vector bundles is the real Grassmannian $G^{\mathbb{R}}_{n,N}$, $N \gg 0$. The cohomology groups of this space have torsion of order 2. Therefore, one needs a separate description of both mod2 and rational cohomology classes. The mod2 cohomology ring of $G^{\mathbb{R}}_{n,N}$ is generated by the classes $w_1, \ldots, w_n$, $w_k \in H^k(G^{\mathbb{R}}_{n,N}, \mathbb{Z}_2)$, called the *Stiefel-Whitney classes*. The theory of these classes for real vector bundles, including the splitting principle and the Whitney formula, is parallel to the theory of Chern classes for complex bundles. Moreover, the complexification principle claims that in order to get some general formula for Stiefel-Whitney classes, in many cases it is sufficient to take the corresponding formula for the Chern classes, to replace c_i by w_i in it, and to reduce the coefficients mod2.

The integer characteristic classes of oriented real vector bundles are the Pontryagin-Euler classes. The *Pontryagin classes* $p_i(E)$ have the grading $4i$, $i \leq \operatorname{rk} E/2$. In the case when $\operatorname{rk} E = 2n$ is even, one has $p_n(E) = e(E)^2$. Otherwise $e(E) = 0$. The Whitney formula for the Pontryagin classes

$$p(E \oplus F) = p(E)\, p(F)$$

holds modulo torsion only.

The non-triviality of the characteristic classes may lead to unavoidability of certain degeneracies of some differential geometry structures. Conversely, the characteristic classes can be computed by studying degeneracies (singularities) of these structures. For instance, the class $w_k(E)$ can be defined as the class Poincaré dual to the locus of points where the given generic $n-k+1$ sections of the bundle E are linearly dependent, where $n = \operatorname{rk} E$.

A similar description exists for the Chern and the Pontryagin classes. There is a very wide range of differential geometry structures whose degenerations can be studied by using characteristic classes. Consider, for example, the singularity loci of generic differential maps $M \to N$. As it was observed by R. Thom, for any fixed singularity type α of map germs $(\mathbb{R}^m, 0) \to (\mathbb{R}^n, 0)$ the Poincaré dual of the corresponding locus in the source manifold M is given by a universal polynomial (called now *Thom polynomial*) in the Stiefel-Whitney classes $w_i(M)$ and $f^* w_j(N)$ determined entirely by the singularity type and independent of the particular choice of M, N, and f. The Thom polynomials of the complex singularities are expressed in a similar way in terms of the Chern classes. The computation of the Thom polynomials for particular types of singularities was initiated in 60-ies and is not yet finished now. The theory of the universal polynomials in characteristic classes has been extended to the study of multisingularities by M. Kazarian [86].

The characteristic classes are useful in the classification of manifolds. By definition, the characteristic classes of a manifold are the characteristic classes of its tangent bundle. This definition suggests that the characteristic classes of a manifold could characterise its smooth structure. It turns out, however, that the Stiefel-Whitney classes are homotopy invariants of the manifold (by the so called Wu formula, they can be recovered from the action of the Steenrod algebra). Moreover, a more complicated theorem due to Novikov claims that the rational Pontryagin classes are also homotopy invariant. However, it is amazing to observe that the integer Pontryagin classes do depend on the smooth structure and for different smooth structures the corresponding Pontryagin classes can differ by a torsion element.

Characteristic numbers. A *characteristic number* is the value of a characteristic class on the fundamental class of the manifold. Thus, there are as many characteristic numbers as the number of monomials in the Stiefel-Whitney (or Pontryagin, or Chern, depending on the situation) classes whose degree is equal to the dimension of the manifold.

The characteristic numbers are used in the cobordism classification of manifolds by the following main observation:
Any two cobordant manifolds have equal values of all their characteristic numbers; conversely, if two manifolds have equal all their characteristic numbers, then they are cobordant.
This holds both in the real case (oriented or not) and in the complex case with the Chern classes instead of the Stiefel-Whitney ones. In the real oriented case, a similar assertion for the Pontryagin characteristic numbers holds modulo torsion (that is, the vanishing of all Pontryagin characteristic numbers implies that the union of several copies of a given oriented manifold is a boundary).

For example, the signature of a $4k$-dimensional manifold (see p. 631) is an invariant of the oriented cobordism class of that manifold. It follows that it can be expressed as a universal combination of the characteristic numbers of that manifold. Thus, studying examples one can determine explicitly all coefficients of this combination. For instance, one computes that the signature of a 4-dimensional manifold M^4 is equal to $\frac{1}{3}p_1$. It follows that $p_1(M^4)$ is divisible by 3 for any compact 4-fold. Similarly, the signature of a 8-fold M^8 is equal to $(7\,p_2 - p_1^2)/45$. Similar formulas with obvious consequences for divisibility properties have been extended to higher dimensions by Hirzebruch.

The theory of characteristic classes of vector bundles can be generalised to the theory of characteristic classes of fibre bundles with a fixed structural Lie group G. The characteristic classes of this theory are the cohomology classes of the classifying space BG. The cases of the complex and the real vector bundles correspond to the Lie groups $\mathrm{U}(n)$ and $\mathrm{O}(n)$, with the respective classifying spaces* $B\mathrm{U}(n) = G^{\mathbb{C}}_{n,\infty}$ and $B\mathrm{O}(n) = G^{\mathbb{R}}_{n,\infty}$.

In the case when the group G is discrete, the corresponding classifying space is the Eilenberg-Maclane space $BG = K(G, 1)$, determined uniquely up to homotopy equivalence by the properties $\pi_1(BG) \simeq G$ and $\pi_k(BG) = 0$ for $k > 1$. For example, the cohomologies of the complement to the swallowtail have the meaning of the characteristic classes of the braid group – the complement of the swallowtail discussed in Ch. 5 (p. 138), is the classifying space of its own fundamental group, and this is true for any space $K(\pi, 1)$.

In symplectic geometry, one constructs also the so-called Lagrangian characteristic classes (of Lagrangian submanifolds) as cohomology classes of the Lagrange Grassmannian $\Lambda_n = \mathrm{U}(n)/\mathrm{O}(n)$. The cohomology groups with coefficients in $\mathbb{Z}_2$ of this space are generated by the Stiefel-Whitney classes that, in this case, satisfy the additional restriction $w_i^2 = 0$. The rational cohomology of Λ_n is generated by certain classes $\mu_1, \mu_5, \mu_9 \ldots,$ where $\mu_{4k+1} \in H^{4k+1}(\Lambda_n)$. The class μ_1 is called the *Arnold-Maslov* characteristic class. The other classes μ_i are related to the Pontryagin classes, but this relationship is not so straightforward. The Lagrangian characteristic classes can be defined for any Lagrangian subvariety in the symplectic space $\mathbb{R}^{2n}$, and more general, for any Lagrangian subvariety of any symplectic cotangent bundle space. They describe the cohomology classes dual to the loci of singularities of the projection of a Lagrangian manifold to the base of the cotangent bundle. It is the Lagrangian version of the Thom theory of characteristic classes related to singularities, due to Vassiliev.

---

*The symbol ∞ means that for each degree k the kth cohomology of the Grassmanian $G_{n,N}$ stabilises from some value of N, which depends on that degree. This "limite" cohomology is called the group of characteristic classes of the given degree.

Chapter 12

Betti numbers of complex surfaces via Morse theory

This chapter is conceived as an application of Morse theory to a fundamental problem: to study the Homology of algebraic complex hypersurfaces in projective space $\mathbb{C}\mathrm{P}^m$.

Our first goal is to solve the following problem for complex surfaces.

Problem. Calculate the five Betti numbers of the smooth complex algebraic surfaces of degree n in the complex projective space $\mathbb{C}\mathrm{P}^3$.

12.1 Two special functions on a special surface

Although the problem is formulated for algebraic complex surfaces in $\mathbb{C}\mathrm{P}^3$, we start from the general case of algebraic hypersurfaces in $\mathbb{C}\mathrm{P}^m$, since most arguments and calculations are independent of m.

It suffices to solve the problem for one example because all such complex hypersurfaces are diffeomorphic as real manifolds – see the discussion of the "Italian principle of algebraic geometry" on page 145.

As example we choose the smooth complex hypersurface M^{m-1} defined in the affine coordinates $(x_1, \ldots, x_m)$ of $\mathbb{C}\mathrm{P}^m$ by the equation (Fig. 12.1)

$$x_1^n + \cdots + x_m^n = 1, \quad \text{that is,}$$

$$M^{m-1} \cap \mathbb{C}^m = \{x \in \mathbb{C}^m : F(x) = 1, \quad \text{where} \;\; F(x) = x_1^n + \ldots + x_m^n\}.$$

The smoothness of M^{m-1} in $\mathbb{C}^m$ follows from the implicit function theorem:

$$\left(\frac{\partial F}{\partial x_j} = 0\right) \iff (x_j = 0) \; ,$$

since at every point of M^{m-1} there exists $x_j \neq 0$ because $\sum x_j^n = 1$. The same reasoning shows the smoothness of M^{m-1} everywhere because the affine equation of M^{m-1} at infinity is of the form $\sum \pm y_j^n = 1$.

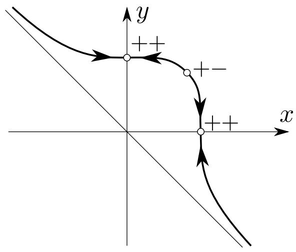


Figure 12.1: The curve $x^3 + y^3 = 1$, three critical points of the function r^2 on it, their signatures and the descent directions.

To study the homology groups, we shall consider the following two real functions:

$$r^2 : \mathbb{C}^m \to \mathbb{R}, \quad r^2 = |x_1|^2 + \cdots + |x_m|^2, \quad \text{and}$$

$$f : \mathbb{C}\mathrm{P}^m \to \mathbb{R}, \quad f(x) = \frac{r^2}{1 + r^2} \ .$$

The function f is smooth in $\mathbb{C}\mathrm{P}^m$ and, in the homogeneous coordinates $X_0 : X_1 : \cdots : X_m$, it has the form

$$f(X) = \frac{|X_1|^2 + \cdots + |X_m|^2}{|X_0|^2 + \cdots + |X_m|^2} \, .$$

It attains the maximal value $f(X) = 1$ at the hyperplane at infinity $\mathbb{C}\mathrm{P}_\infty^{m-1} = \{X_0 = 0\} \approx \mathbb{C}\mathrm{P}^m \smallsetminus \mathbb{C}^m$.

We shall study the Morse complex of the real function f on the complex hypersurface M, considering it as real manifold dimension $2m - 2$.

The special choice of F and f provides a finite symmetry group of order $m!n^m$, formed by the $m!$ permutations of the coordinates x_j and by the multiplications of each of them (independently of each other) by the n roots of degree n of 1. Since in this example the Morse complex of f on M^{m-1} is invariant under this group action, the calculations are $m!n^m$ times shorter than in the generic cases. The following proposition helps to simplify calculations.

Proposition. *In the finite part of M the functions $r^2_{|M}$ and $f_{|M}$ have the same critical points with the same respective Morse indices.*

Proof. In the finite part $\mathbb{C}^m$ of $\mathbb{C}\mathrm{P}^m$, the critical points of $f_{|M}$ coincide with the critical points of $r^2_{|M}$, since

$$f = 1 - \frac{1}{1+r^2}\,, \qquad df = +\frac{dr^2}{(1+r^2)^2}\,,$$

and thus $(df = 0) \Leftrightarrow (dr^2 = 0)$.

Moreover,

$$d^2 f = \frac{d^2(r^2)}{(1+r^2)^2} - \frac{2\big(d(r^2)\big)^2}{(1+r^2)^3}\,.$$

Hence the signatures of the quadratic forms $d^2(r^2)$ and $d^2 f$ on the tangent plane at the critical point, at which $d(r^2) = 0$, are equal.

The ordering of the critical levels of the function $f_{|M}$ is also the same as for the function $r^2_{|M}$ because the dependence $f = r^2/(1+r^2) = 1 - 1/(1+r^2)$ is monotone. □

Therefore to describe the critical points of the function $f_{|M}$ in the finite part of M it suffices to study the critical points of $r^2_{|M}$.

12.2 Critical points of the function $r^2_{|M}$

The following Lemma (proved along this section) describes all critical points of the restriction of r^2 to $M = \{x \in \mathbb{C}^m : x_1^n + \cdots + x_m^n = 1\}$, for $n > 2$.

Given $k \in \mathbb{N}$, we note $a_k := 1/k^{1/n}$ the positive real n-tic root of $1/k$.

Main Lemma A. *For each subset of k elements $J_k \subset \{1, 2, \ldots, m\}$, with $1 \leqslant k \leqslant m$, the n^k points of $\mathbb{C}^m$ with coordinates $x_j = 0$ for $j \notin J_k$ and*

$$x_j = a_k\, e^{i\vartheta_j} \quad \textit{for} \;\; j \in J_k\,, \quad \textit{where} \;\; e^{in\vartheta_j} = 1\,,$$

are non-degenerate critical points of the restriction of r^2 to the hypersurface $M = \{x \in \mathbb{C}^m : \sum x_j^n = 1\}$, $r^2_{|M}$. This function has no other critical point.

B. *The Morse index of these n^k critical points is equal to $k-1$.*

Remark. Taking all possible subsets J_k for a fixed k, we get $c_{k-1} := \binom{m}{k} n^k$ critical points of index $k-1$. All of them have the critical value $r^2 = k^{1-2/n}$. Therefore the function $r^2_{|M}$ has $\sum_{k=1}^m \binom{m}{k} n^k = (n+1)^m - 1$ critical points.

Example. For $m = n = 3$, the function $r^2_{|M}$ has $4^3 - 1 = 63$ critical points:

$$\begin{array}{rl} 9 = 3n & \text{critical points of signature } (+++\,+),\ r^2 = 1; \\ 27 = 3n^2 & \text{critical points of signature } (-++\,+),\ r^2 = \sqrt[3]{2}; \\ 27 = n^3 & \text{critical points of signature } (--++),\ r^2 = \sqrt[3]{3}; \end{array}$$

the simplest examples being the 3 real points $(a_1, 0, 0)$, $(a_2, a_2, 0)$, (a_3, a_3, a_3). All other critical points are obtained from these 3 points by permuting the coordinates and multiplying them by the cubic roots of 1.

Remark. The inequality "(Morse index) $\leqslant m-1$" for the critical points of r^2 on M (implied by the Main Lemma) is a general fact of "Stein theory" for the holomorphic hypersurfaces in $\mathbb{C}^m$. In fact, our reasoning below prove the following fact: The Morse index of a critical point of the function r^2 on a holomorphic hypersurface in $\mathbb{C}^m$ depends on the principal curvatures in the following way. A complex direction with zero curvature (or small curvature) provides 2 positive squares $(++)$, while a complex direction of large curvature provides the signs $(+-)$ to the signature. Here the curvature is "large" if the curvature radius is smaller than the distance of the critical point to the origin.

In order to prove Main Lemma A, we need the following

Proposition. *At a critical point z_0 of $r^2_{|M} : M \to \mathbb{R}$, the tangent hyperplane to the complex hypersurface M is Hermitian orthogonal to the vector z_0.*

Proof. Choose a Hermitian coordinate system at the critical point z_0 by taking the first coordinate axis t directed along the vector z_0, with the other axes $(w_1, \ldots, w_{m-1})$ being Hermitian orthogonal to it (Fig. 12.2).

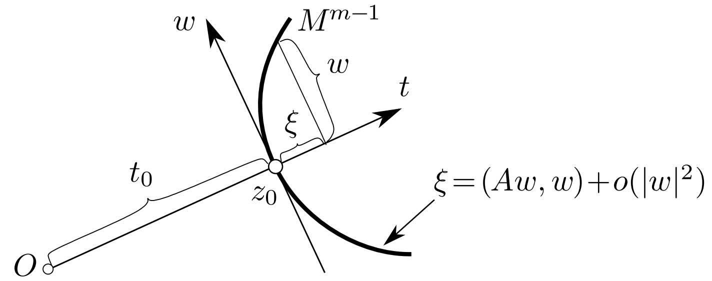


Figure 12.2: The hypersurface M near a critical point z_0 of the function $r^2_{|M}$

We shall prove that, in this coordinate system, the hyperplane $t = t_0$ is tangent to M at the critical point $z_0 = (t_0, 0)$. Indeed, in Euclidean complex scalar product notation, $(\,\cdot\,,\cdot\,)$, the value of r^2 at $z = z_0 + (\xi, w)$ is

$$\begin{aligned} r^2 &= (t+w)(\overline{t}+\overline{w}) \\ &= (t_0+\xi, \overline{t_0}+\overline{\xi}) + |w|^2 \\ &= |t_0|^2 + 2\mathrm{Re}\,(\overline{t_0}\xi) + |\xi|^2 + |w|^2\,. \end{aligned}$$

Therefore, $\mathrm{Re}\,(\overline{t_0}\xi) = 0$ for any tangent vector (ξ, w) of the hypersurface M because $z_0 = (t_0, 0)$ is a critical point of the restriction of r^2 to M.

Since $t_0 \neq 0$, the condition "$\mathrm{Re}\,(\overline{t_0}\xi) = 0$ for all the tangent vectors" implies that $\xi = 0$ for all these vectors. Therefore, at $(t_0, 0)$ all tangent vectors to the hypersurface are Hermitian orthogonal to the vector z_0. □

Proof of Main Lemma A (**Locating the critical points**). Since our hypersurface M is defined by the equation $F(x) = 1$, where $F(x) = x_1^n + \ldots + x_m^n$, its tangent hyperplane is determined by the first differential:

$$\frac{\partial F}{\partial x_1}\,dx_1 + \cdots + \frac{\partial F}{\partial x_m}\,dx_m = 0\,. \tag{1}$$

The Hermitian orthogonality of this hyperplane to the vector z_0, proved above, implies that

$$\overline{x}_1\,dx_1 + \cdots + \overline{x}_m\,dx_m = 0$$

for every tangent vector that verifies the condition (1).

Therefore, we obtain the conditional extremum equations

$$\overline{x}_1 = \lambda\,\frac{\partial F}{\partial x_1}\;,\;\ldots\;,\;\overline{x}_m = \lambda\,\frac{\partial F}{\partial x_m} \tag{2}$$

where λ is a complex constant. To find λ, we apply to F (which is homogeneous of degree n) the Euler identity for homogeneous functions:

$$\sum(\overline{x}_j x_j) = \lambda \sum\left(x_j \frac{\partial F}{\partial x_j}\right) = \lambda n F = \lambda n\ .$$

Hence the constant $\lambda = r^2/n$ is a positive real number (independent of j).

Writing equations (2) for our $F = \sum x_j^n$, we obtain

$$\overline{x}_j = \lambda n x_j^{n-1}\ . \tag{3}$$

Now, in order to solve equations (3), we use the trigonometric expression ($\rho_j = |x_j|,\ x_j = \rho_j e^{i\vartheta_j}$), obtaining the equalities:

$$\rho_j = \lambda n \rho_j^{n-1}\,, \quad e^{-i\vartheta_j} = e^{i(n-1)\vartheta_j}\,, \tag{4}$$

which imply for each fixed j either $\rho_j = 0$ or $\rho_j^{n-2} = 1/\lambda n$ with $e^{in\vartheta_j} = 1$.

Let k be the number of those coordinates x_j different from zero at the considered critical point. The modulus ρ_j of these k coordinates has the same value ρ, which we obtain from the equalities $F(x) = k\rho^n = 1$, that is,

$$\rho = 1/k^{1/n} =: a_k\,. \qquad \square$$

Exercise. Show that the statement of Main Lemma does not hold for $n = 1$ and $n = 2$.

Hint. First: find the step of the proof which does not apply for $n = 1$ (resp. for $n = 2$). Next: adapt it to find the critical points of $r^2_{|M}$ for $n = 1$ (resp. for $n = 2$).

12.2.1 Computing the Morse indices

We start by reminding an elementary property of complex quadradic forms, useful to study complex submanifolds of $\mathbb{C}^\ell$ as real submanifolds of $\mathbb{R}^{2\ell}$: If Q is a complex quadratic form on $\mathbb{C}^\ell$, its real part $\Phi = \operatorname{Re} Q$ is a real quadratic form on $\mathbb{R}^{2\ell} = \mathbb{C}^\ell$ which satisfies the condition $\Phi(iw) = -\Phi(w)$. Therefore, in its normal diagonal expression, it can be written as

$$\Phi = \sum_{j=1}^{\ell} c_j(p_j^2 - q_j^2) \quad (\text{all } c_j \geqslant 0) \tag{5}$$

in some Euclidean orthonormal coordinate system. This implies that *The real part of a complex quadratic form, $\Phi = \operatorname{Re} Q$, has signature zero* (i.e., in its diagonal normal form (5) the number of positive terms equals the number of negative terms).

To study the signature of $r^2_{|M}$, we use the above formula (of p. 475) at the point $(t_0 + \xi, w)$, in the adapted coordinates (t, w):

$$r^2 = |t_0|^2 + 2\operatorname{Re}(\overline{t_0}, \xi) + |\xi|^2 + |w|^2\,.$$

In these coordinates (Fig. 12.2), the local equation of the hypersurface M has the Taylor form

$$\xi = (Aw, w) + o(|w|^2)$$

(the first differential vanishes, as we have proved). Therefore,

$$r^2 = r_0^2 + |w|^2 + 2\,\mathrm{Re}\,\big(t_0(Aw, w)\big) + o(|w|^2)\,.$$

The real quadratic form $2\,\mathrm{Re}\,\big(t_0(Aw, w)\big)$ of variable w has signature 0 because it changes the sign when the vector w is multiplied by i.

Therefore *the Morse index of a critical point of the function* $r^2_{|M} : M \to \mathbb{R}$ *of* $2(m-1)$ *real variables cannot exceed the number* $m-1$ *of the complex coordinates* w*, whatever be the holomorphic hypersurface* M^{m-1} *in* $\mathbb{C}^m$.

We shall prove that for our special hypersurface M the Morse index is equal to $k-1$ at every critical point of $r^2_{|M}$ having exactly k coordinates different from 0. Geometrically it means that the corresponding $k-1$ curvature radii are smaller than the distance to 0.

To make it easy, we consider the real critical point z_0 with coordinates

$$x_1 = \cdots = x_k = a\ ,\ x_{k+1} = \cdots = x_m = 0 \qquad \text{(where } a = a_k)\ . \tag{6}$$

All other critical points which have exactly k nonzero coordinates can be sent to this one by the symmetries that preserve F and r^2. Therefore all these points have the same signature and the same Morse index.

In the special case (6), we choose the coordinates $(x_{k+1}, \ldots, x_m)$ to be a part of the w-coordinates because $\partial F/\partial x_j = 0$ for $x_j = 0$. Observe that if $n > 2$ we also have for these coordinates $\partial^2 F/\partial x_j^2 = 0$ for $x_j = 0$.

The remaining $k-1$ coordinate axes w are directed along the subspace $dx_1 + \cdots + dx_k = 0$ of $\mathbb{R}^k$, orthogonal to the vector (6).

The quadratic form $2t_0(Aw, w)$ of the variable vector w is invariant under the permutations of the first k coordinate axes (because z_0 is fixed by these permutations and the functions F and r^2 are invariant...). This implies the equality $2t_0(Aw, w) = cw^2$ in the real orthogonal space $\mathbb{R}^{k-1}$. We choose some orthogonal basis in this real space and denote by $w'_1, \ldots, w'_{k-1}$ the corresponding Cartesian coordinates in $\mathbb{R}^{k-1}$. Denoting the same way these complex linear functions in $\mathbb{C}^{k-1}$, we get there

$$2t_0(Aw, w) = c\,\big((w'_1)^2 + \cdots + (w'_{k-1})^2\big)\ , \tag{7}$$

where c is a complex constant. Our goal is to prove the following Lemma.

Lemma. *In our example, the constant* c *of* (7) *is the real number* $1-n$.

Proof. Let $x = z_0 + \eta$ be a perturbation in M of our critical point z_0 by a small vector η of $\mathbb{R}^k$, with coordinates $x_j = a + \eta_j$ $(1 \leqslant j \leqslant k)$, $x_j = 0$ $(j > k)$. Since the value of F at x is given by the formula

$$F(x) = ka^n + na^{n-1}\sum_{j=1}^{k}\eta_j + \frac{n(n-1)}{2}a^{n-2}\sum_{j=1}^{k}\eta_j^2 + o(|\eta|^2)$$

and $F(x) = 1 = ka^n$ (because x and z_0 belong to M), the local equation of the real surface $M \cap \mathbb{R}^k$ at z_0 is

$$2a\sum_{j=1}^{k}\eta_j = -(n-1)\sum_{j=1}^{k}\eta_j^2 + o(|\eta|^2)\,. \tag{8}$$

We shall rewrite (8) in the adapted coordinates (t, w), to get the local expression of $M \cap \mathbb{R}^k$ as the graph of a function $\xi = \xi(w)$ (see Fig. 12.2).

The component ξ of the vector $\eta = (\eta_1, \ldots, \eta_k, 0, \ldots, 0)$ along the t-axis is the sum of the projections of the components η_j of η on this t-axis: $\xi = \sum_{j=1}^{k}\eta_j/\sqrt{k}$. Now, since $|z_0| = t_0 = a\sqrt{k}$, the left hand side of (8) is equal to $2t_0\xi$.

Moreover, since $\sum_{j=1}^{k}\eta_j^2 = w^2 + \xi^2$ for the real vector w orthogonal to $z_0 = (a, \ldots, a, 0, \ldots, 0)$, equation (8) becomes $2t_0\xi = (1-n)(w^2 + \xi^2) + o(|w|^2)$, equivalent to

$$2t_0\xi = (1-n)w^2 + o(|w|^2)\,,$$

which provides the quadratic term of the Taylor series of $\xi(w)$ along M and proves that the constant c of (7) is the real negative number $1 - n$. □

If $n > 2$, then $c < -1$, which means that the curvature of the submanifold M at our critical point is “large” (in the sense of page 474).

Proposition. *If a critical point of r^2, restricted to the hypersurface of degree $n > 2$*

$$M = \left\{x \in \mathbb{C}^m : \sum_{j=1}^{m} x_j^n = 1\right\},$$

has exactly k nonzero coordinates x_j, then its Morse index is equal to $k - 1$.

Proof. In the coordinates w'_j introduced above, the expression of r^2 is

$$r^2 = r_0^2 + \left(\sum_{j=1}^{k-1} \left(c\,\mathrm{Re}\,(w'_j)^2 + |w'^2_j|\right) + \sum_{j>k} |x_j|^2 + o(|x - z_0|^2) \right),$$

where $c = 1 - n < -1$ because $n > 2$. Hence, for $w'_j = p + iq$ we get

$$\begin{aligned}\left(c\,\mathrm{Re}\,(w'_j)^2 + |w'^2_j|\right) &= c(p^2 - q^2) + (p^2 + q^2)\\ &= q^2(1 - c) + p^2(1 + c),\end{aligned}$$

in which the coefficient of q^2 is positive and that of p^2 negative because $c = 1 - n < -1$. Hence, at a critical point z_0 with k nonzero coordinates, the negative inertia index of the quadratic form $d^2(r^2)$ is equal to $k - 1$. □

We have thus proved Main Lemma B (for hypersurfaces of degree $n > 2$).

12.3 Euler caracteristic of a hypersurface in $\mathbb{C}\mathrm{P}^m$

To compute the Euler caracteristic of a complex hypersurface (Theorem 1 below), we need the following two corollaries of the Main Lemma. They describe the critical points of a perturbation at infinity, $\widetilde{f}$, of our function f.

Corollary 1. *The real function* $f = \dfrac{r^2}{1 + r^2}$ *on the* $(2m-2)$*-dimensional real manifold* $M \subset \mathbb{C}\mathrm{P}^m$ *given by the equation* $\sum_{j=1}^m x_j^n = 1$ *has* $c_{k-1} = \binom{m}{k} n^k$ *"finite" non-degenerate critical points of index* $k - 1$ *(where* $1 \leqslant k \leqslant m - 1$*).*

The function f *attains the maximal value* $f = 1$ *on the infinitely far submanifold*

$$\widetilde{M}^{m-2} = M^{m-1} \cap \mathbb{C}\mathrm{P}^{m-1}_{\infty},$$

*which is Bott non degenerate**: *The second differential of* f *is negative-definite on the 2-planes transverse to the submanifold* $\widetilde{M}^{m-2}$ *in* M^{m-1}*.*

The function f *has no other critical points on* M^{m-1}*.*

---

*Classical Morse theory deals only with functions all of whose critical points are non-degenerate; in particular, the critical points must all be isolated points. In many situations, however, the critical points form submanifolds of the manifold where the function is defined. A critical manifold B is said to be *Bott non degenerate* if every critical point of it becomes of Morse type when the function is restricted to a transverse slice to B.

Proof. The fist part of Corollary 1, concerning the finite part of M, follows from Main Lemma and the Proposition of p. 472.

At infinity we use other affine coordinates, say,

$$x_1 = 1/X_1 \ , \ x_2 = X_2/X_1 \ , \ \dots \ , \ x_m = X_m/X_1 \ ,$$

for which the equation $F(x) = 1$ takes the form $1 + \sum_{j=1}^m X_j^n = X_1^n$.

In these new coordinates, the function f takes the form

$$f = \frac{\sum_{j=2}^m |X_j|^2}{\sum_{j=1}^m |X_j|^2} = 1 - \frac{|X_1|^2}{\sum_{j=1}^m |X_j|^2} \ .$$

Hence along the submanifold at infinity $\widetilde{M}^{m-2} = M^{m-1} \cap \mathbb{C}\mathrm{P}_\infty^{m-1}$ $(X_1 = 0)$ the function f takes the value $1 - 0 = 1$.

Along a 2-plane transverse to the hypersurface $\widetilde{M}^{m-2}$ in M^{m-1}

$$f = 1 - \frac{|X_1|^2}{K + |X_1|^2} \ ,$$

where $K = \sum_{j=2}^m |X_j^2| > 0$. Therefore $d^2 f < 0$, proving Corollary 1.

It is sufficient to consider just one (any) transverse slice, e.g. the one used in the proof. The footnote on the previous page, refers correctly to "a transverse slice" meaning that the choice of this slice does not matter. Indeed, the tangent space to the critical manifold is the kernel of the second differential, i.e. $d^2 f$ is well defined on the quotient space. And the tangent space to any transversal projects to this quotient isomorphically. □

Corollary 2. *On the $(2m-2)$-dimensional real manifold $M \subset \mathbb{C}\mathrm{P}^m$, where $\sum_{j=1}^m x_j^n = 1$, there exists a smooth real function $\widetilde{f}$ whose number of non degenerate critical points of index k (with $0 \leqslant k \leqslant 2m-2$) is given by*

$$\widetilde{c}_k(m) = c_k(m) + c_{k-2}(m-1) + c_{k-4}(m-2) + \dots \ , \tag{9}$$

where $c_k(\ell) = \binom{\ell}{k+1} n^{k+1}$ for $1 \leqslant k+1 \leqslant \ell$.
The function $\widetilde{f}$ has no other critical point.

Proof. Deforming slightly our initial Bott function f of corollary 1 at infinity, we obtain a function $\widetilde{f}$. Writing $\widetilde{c}_{k-2}(m-1)$ for the number of critical points of index $k-2$ on $\widetilde{M}^{m-2}$, we have that the number of critical points of $\widetilde{f}$ is given by

$$\widetilde{c}_k(m) = c_k(m) + \widetilde{c}_{k-2}(m-1) \tag{10}$$

because a critical point of index $k-2$ of the perturbed function $\widetilde{f}$ on the submanifold at infinity $\widetilde{M}^{m-2} = M^{m-1} \cap \mathbb{C}\mathrm{P}_\infty^{m-1}$ provides a critical point of index k for this perturbed function on M^{m-1}.

To find $\widetilde{c}_{k-2}(m-1)$ observe that the manifold $\widetilde{M}^{m-2}$ is diffeomorphic to

$$\left\{ z \in \mathbb{C}\mathrm{P}^{m-1} : \sum_{j=1}^{m-1} z_j^n = 1 \right\}$$

because the standard affine coordinates at infinity describe $\widetilde{M}^{m-2}$ as the manifold defined by the equation $\sum_{j=1}^{m-1} (\pm 1) z_j^n = 1$. Its diffeomorphism to the above normal form is provided by the multiplications of some of the coordinates by $\sqrt[n]{-1}$.

Therefore, for $\widetilde{c}_{k-2}(m-1)$ we get the formula

$$\widetilde{c}_{k-2}(m-1) = c_{k-2}(m-1) + \widetilde{c}_{k-4}(m-2) ,$$

which, with formula (10), provides expression (9) by induction on m. □

Exercise. Check that for $m = 3$ the numbers $\widetilde{c}_k(3)$ of critical points of index k of the smooth function $\widetilde{f}$ on the complex algebraic surface $M^2 \subset \mathbb{C}\mathrm{P}^3$ defined by the equation $x^n + y^n + z^n = 1$, are the following:

$$\widetilde{c}_0 = 3n , \quad \widetilde{c}_1 = 3n^2 , \quad \widetilde{c}_2 = n^3 + 2n , \quad \widetilde{c}_3 = n^2 , \quad \widetilde{c}_4 = n .$$

Hint: We have $c_k(\ell) = 0$ for $k \geqslant \ell$ because the index is at most $\ell - 1$ (see formula (9)).

We get the following expression for the Euler characteristic of M^2:

$$\chi = \sum_{k=0}^{2m-2=4} (-1)^k \, \widetilde{c}_k(m) = n^3 - 4n^2 + 6n . \tag{11}$$

Theorem 1. *The Euler characteristic of any smooth complex hypersurface M of degree n in projective space $\mathbb{C}\mathrm{P}^m$ is*

$$\chi(M) = m + \frac{n-1}{n} + \frac{(1-n)^{m+1}}{n} . \tag{12}$$

Example. For $m = 2$ the expression (12) takes the form

$$\begin{aligned} \chi &= 2 + 1 - \frac{1}{n} + \frac{1 - 3n + 3n^2 - n^3}{n} \\ &= 3n - n^2 \\ &= 2 - 2g \end{aligned}$$

because $2g = (n-1)(n-2) = n^2 - 3n + 2$ (see p. 148-149).

Exercise. Check that for $m=3$ we get again formula (11).

Proof of Theorem 1. Fix n and consider the Euler characteristic of M^{m-1} as a function $\chi(m)$. According to formla (9) of Corollary 2 and to the additivity formula for the Euler characteristic, $\chi = \sum_{k=0}^{2m-2}(-1)^k \widetilde{c}_k(m)$, we have

$$\chi(m) = \sum_{j\leqslant m} S(j) \ , \text{ where } S(j) = \sum_{k+1=1}^{j} (-1)^k c_k(j)$$

($S(j)$ is the Euler characteristic of the finite part of M^{j-1} in $\mathbb{C}\mathrm{P}^j$).

The value of $S(j)$ is obtained from the Morse complex of $r^2_{|M}$:

$$-S(j) = \sum_{k+1=1}^{j} \binom{j}{k+1} (-n)^{k+1} = (1-n)^j - 1 \ .$$

Theorem 1 follows by the summation of a geometric progression :

$$\chi(m) = \sum_{j=1}^{m} \left(1-(1-n)^j\right) = m - (1-n)\frac{1-(1-n)^m}{1-(1-n)} =$$

$$= m + \frac{n-1}{n} + \frac{(1-n)^{m+1}}{n} \ .$$

□

12.4 Computing the first Betti number $b_1(M^{m-1})$

The information we have already on the Morse complex of the functions r^2 and f on M and on their symmetries (the boundary operation commuting with the action of the symmetries) can be used to calculate the homology. For instance, if a "finite" critical point lives in a coordinate subspace where some of the coordinates x_j vanish, then all the descending gradient lines of r^2 (or of f) remain in this coordinate subspace. This follows, for instance, from the above reduction of the second differential of r^2 to the normal form.

Hence, the Morse complex differential defines the boundary of the k-cell that corresponds to a critical point of index k : It consists of the $(k-1)$-cells that correspond to the critical points in the same subspace (namely, they belong to the coordinate hyperplanes of this subspace). See below Fig. 12.4.

Example. The mn minima points of index $k = 0$ form m regular n-gons in the m complex coordinate lines: They are the roots of degree n of 1.

The $\binom{m}{2}n^2$ critical points of index 1 are the symmetry images of the point (a_2, a_2) of the plane curve $x_1^n + x_2^n = 1$. The descending 1-manifold, which is real in this example, leads to the two minima points $(a_1, 0)$ and $(0, a_1)$. Therefore, the boundary operator is

$$\partial[(a_2, a_2)] = [(a_1, 0)] - [(0, a_1)]\,,$$

where $[\,\cdot\,]$ means the Morse complex cell generated by the critical point $(\cdot)$ and the signs depend on the arbitrary chosen orientations of the cells.

The $\binom{m}{2}n^2$ critical points of index 1 provide the 1-dimensional complex of mn vertices, whose edges connect every vertex of each n-gon to every vertex of every other n-gon – in Fig. 12.3 the n-gons are horizontal.

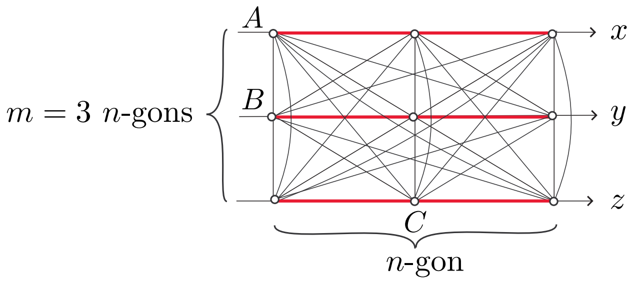


Figure 12.3: The 1-dimensional part of the Morse complex of the function r^2 on the surface $x^3 + y^3 + z^3 = 1$.

Triangular cycles. The 1-dimensional cycles of the Morse complex are just the cycles of this 1-dimensional complex: they are the closed chains whose edges join vertices of distinct n-gons; thus the sides and diagonals of the n-gons do not belong to the complex. There exist triangular closed cycles of this type – like the cycle ABC shown in Fig. 12.3.

Theorem. *The first Betti number of any hypersurface $M^{m-1} \subset \mathbb{C}\mathrm{P}^m$ of degree n vanishes for $m > 2$, while for $m = 2$ it is $b_1(M^1) = (n-1)(n-2)$.*

Proof. For $m = 2$ we have proved on p. 481 that $b_1(M^1) = 2g = (n-1)(n-2)$. The fact that for $m > 2$ all 1-dimensional cycles of our Morse complex are homologous to zero is a consequence of the following two lemmas. $\square$

Lemma 1. *If $m > 2$, then every closed cycle of the Morse complex may be decomposed into a sum of the above triangular cycles.*

Proof. Indeed, a cycle $A_1A_2A_3\dots A_p$ (with $A_p = A_1$) visiting more than two n-gons, should have a vertex, say A_q, whose neighbours in the cycle are on different n-gons – otherwise all the vertices A_{2k} should belong to one n-gon, and all the A_{2k+1} to other n-gon.

Replacing $A_{q-1}A_qA_{q+1}$ with $A_{q-1}A_{q+1}$, we represent the initial cycle as the sum of the triangular cycle $(A_{q-1}A_qA_{q+1})$ and a cycle of length $p-1$.

In this way we decompose any cycle by the diagonals in a combination of triangular cycles and of cycles visiting only two n-gons.

If $m > 2$, these last cycles are also representable as combinations of triangles. Namely, let $A_1B_1A_2B_2\dots A_pB_p(A_1)$ be such a cycle, where the vertices A_i belong to one n-gon and the vertices B_i to the other.

Construct the chain formed by the triangles

$$CA_1B_1\ ,\ CB_1A_2\ ,\ CA_2B_2\ ,\ \dots\ ,\ CB_pA_1\ ,$$

where C belongs to a third n-gon. The boundary of this chain is the given cycle $A_1B_1, A_2B_2\dots A_pB_p$. □

Lemma 2. *For $m > 2$ every triangular chain is homologous to 0.*

Proof. The three vertices z_A, z_B, z_C, belonging to three different n-gons, correspond to the three respective axis x_A, x_B, x_C.

There exists n^3 critical points of index 2 with equal values of these three $|x_j|$ and with arbitrary values of $\sqrt[n]{1}$. One of these critical points, $z_{A,B,C}$, corresponds to the triple (z_A, z_B, z_C). The boundary of the corresponding 2-cell $[z_{A,B,C}]$ is $[z_{AB}] + [z_{BC}] + [z_{CA}]$ (Fig. 12.4), proving the triviality of the homology class of the triangle. □

12.4.1 Digression on the vanishing $b_1(M) = 0$ for $m > 2$

Our proof was based on two geometrical facts:

1. *Every 1-dimensional cycle of the complex hypersurface M^{m-1} is homologous to a 1-dimensional cycle of the section M^1 of $M^{m-1} \subset \mathbb{C}\mathrm{P}^m$ by a 2-plane $\mathbb{C}\mathrm{P}^2$;*
2. *Every 1-dimensional cycle γ of the algebraic complex curve M^1 is homologous to zero in any surface M^2 of which M^1 is a plane section.*

Both facts have interesting extensions and explanations.

1. The reduction to the section follows from a more general fact:

Proposition. *The homology to a cycle in the section may be replaced by the homotopy to that cycle.*

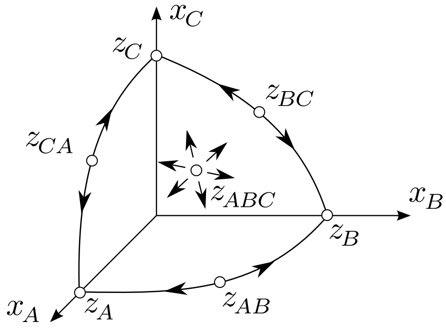


Figure 12.4: The boundary of the descending 2-cell of the critical point z_{ABC} consists of the descending cells of the critical points z_{AB}, z_{BC} and z_{CA}.

Proof. Since the real codimension of the infinite part of M^{m-1} is equal to 2, we may deform the real 1-dimensional curve γ to avoid this infinite part and hence it suffices to prove the homotopy in the affine part $\mathbb{C}^m \subset \mathbb{C}\mathrm{P}^m$.

Let $\pi : \mathbb{C}^m \to \mathbb{C}^{m-1}$ be a generic projection. Its restriction to the hypersurface M^{m-1} is an n-fold covering of $\mathbb{C}\mathrm{P}^{m-1}$ outside some "discriminant hypersurface" $\Sigma^{m-2} \subset \mathbb{C}^{m-1}$. This complex algebraic hypersurface has real codimension 2 in $\mathbb{C}^{m-1}$. Therefore we can suppose that the curve $\pi\gamma$ does not intersect Σ.

According to the Zariski Theorem (of p. 142), the curve $\pi\gamma$ is homotopic to a curve $\widetilde{\gamma}$ in a (generic) one-dimensional affine subspace $\mathbb{C} \subset \mathbb{C}^{m-1}$.

This homotopy inside $\mathbb{C}^{m-1} \setminus \Sigma$ is liftable to M^{m-1}, since the restriction of the projection π to the corresponding part of M^{m-1} is a covering.

The lifted curve $\widetilde{\gamma}$ belongs to the affine plane $\mathbb{C}^2 = \pi^{-1}\mathbb{C}$. This provides the homotopy of the initial curve γ inside M^{m-1} to a curve in the section $\mathbb{C}^2 \cap M^{m-1}$. $\square$

2. The triviality of the cycles in the surface follows from the vanishing cycles theory:

The intersection of the surface $M^2 \subset \mathbb{C}\mathrm{P}^3$ with a subspace $\mathbb{C}\mathrm{P}^2 \subset \mathbb{C}\mathrm{P}^3$ degenerates when the subspace becomes tangent to the surface. At this moment some cycle γ (of real dimension 1) on the algebraic curve $C = M^2 \cap \mathbb{C}\mathrm{P}^2$ vanishes (Fig. 12.5) – like the circle $x^2 + y^2 = c$, along which the surface $z = x^2 + y^2$ intersects the plane $z = c$, degenerates for $c \to 0$.

Varying the families of plane sections, one can prove the vanishing of many cycles of the plane section C. To see that the vanishing cycles generate all the one-dimensional cycles, one have to provide geometrically the basis of the $2g = (n-1)(n-2)$ independent 1-cycles. For instance, there exist real M-curves* of degree n, whose $g+1$ ovals generate a Lagrangian subspace of dimension g: Since the curve divides the plane into two parts and the boundary of one of these parts is formed by the ovals, the $g+1$ ovals are not independent in the homology of dimension one.

*A real *M-curve* of degree n is a curve having the maximal possible number of ovals that a curve of degree n can have: $\frac{(n-1)(n-2)}{2} + 1$.

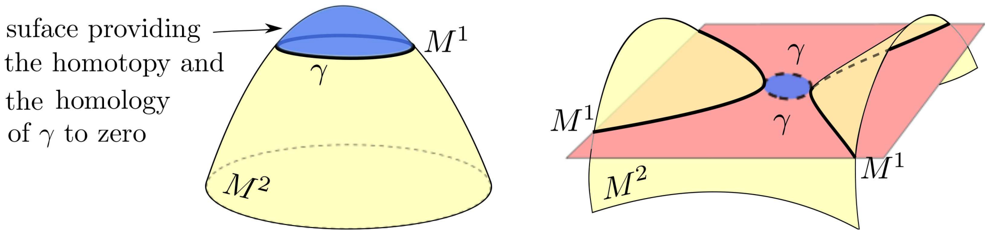


Figure 12.5: The real and imaginary vanishing cycles $\gamma \subset M^1$ are homologous to zero on the complex surface M^2.

Remark. The word "Lagrangian" above, coming from symplectic geometry, means "ovals non intersecting each other": Since the intersection form on the $2g$-dimensional space $H_1(M^2)$ is skew-symmetric, it can be thought as a symplectic form (see Ch. 16). Non-intersecting ovals generate an isotropic subspace.

In the real case, the convex part of the real surface provides small ovals whose homotopy and homology to zero are provided by the real surface bounded by the oval.

At a hyperbolic point, the tangent real plane provides a (less evident) imaginary vanishing cycle, which connects two points of the hyperbola, ready to degenerate at a crossing point of two transverse branches. These vanishing cycles, both of elliptic and hyperbolic types, provide the basis of the 1-dimensional homology group of the complex plane section, for a suitable choice of an algebraic surface with many bumps, whose equation may be written and studied explicitly, but we leave this to the reader.

The extensions of these geometrical studies to the higher-dimensional cases are still waiting the courageous researches. Even the explicit combinatorial description of the differentials of the above Morse complex, taking the large symmetry group into account, should be useful: It has a similar behaviour to that of the so-called Koszul complex*, with unusual coefficients.

12.5 Betti numbers of complex surfaces in $\mathbb{C}\mathrm{P}^3$

In the case of surfaces ($m = 3$) our complex provides all Betti numbers, solving our initial problem. The general formulation is solved in section 12.6.

* **Koszul complex**: Consider the space of germs of holomorphic differential forms at the origin (like $f_1 dx_1 \wedge dx_3$), and choose a function f with a critical point at the origin. Using the differential of f, one defines the "differential map": $\delta : \omega \mapsto \omega \wedge df$, which clearly satisfies $\delta^2 = 0$ and hence it defines a complex, called the *Kozul complex*.

This complex is almost exact, except for the maximal degree, $\deg(\omega^{n-1} \wedge df) = n$: Take $* \, dx_1 \wedge \ldots \wedge dx_n$ where $*$ is a linear combination of the partial derivatives of f.

Passing to the homology, one gets a space that has the dimension of the local algebra.

Theorem. *The Betti numbers of any smooth complex algebraic surface of degree n, $M^2 \subset \mathbb{C}\mathrm{P}^3$, are equal to*

$$b_0 = 1\ ,\ b_1 = 0\ ,\ b_2 = n^3 - 4n^2 + 6n - 2\ ,\ b_3 = 0\ ,\ b_4 = 1\ .$$

Proof. The connexity of the compact complex surface M^2 provides the evident relations $b_0 = 1$, $b_4 = 1$.

The vanishing of the 1-dimensional homology, proved above, implies the relations $b_1 = b_3 = 0$, since b_3 is equal to b_1 by the Poincaré duality.

The Euler characteristic is related to the dimensions $\widetilde{c}_i$ of the chain spaces by the Euler–Poincaré identity (see Fig. 12.6):

$$\chi = \sum(-1)^i \widetilde{c}_i = \sum(-1)^i b_i\,.$$

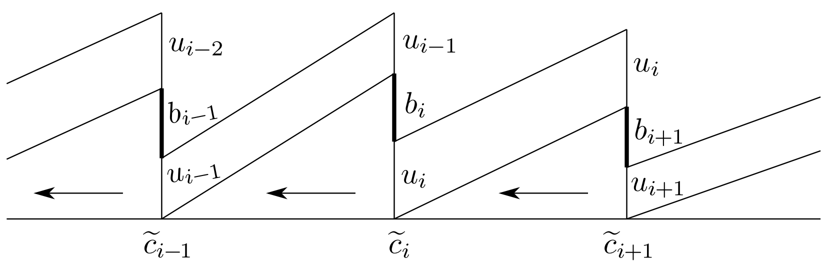


Figure 12.6: The Euler-Poincaré identity $\sum(-1)^i\widetilde{c}_i = \sum(-1)^i b_i$.

It follows from the fact that the image of a differential provides equal positive contributions u_i to two neighbouring dimensions $\widetilde{c}_i$, $\widetilde{c}_{i+1}$, cancelling in the alternate sum:

$$\widetilde{c}_i = u_i + b_i + u_{i-1}\,.$$

If you wish, you can write explicitly

$$u_i = \dim\left(Im(\partial : C_{i+1} \to C_i)\right) = \dim\left(C_{i+1}/\mathrm{Ker}\left(\partial : C_{i+1} \to C_i\right)\right)$$

In other terms, one can introduce a basis in C_i consisting of $c_i = u_i + b_i + u_{i-1}$ elements such that $\partial : C_i \to C_{i-1}$ maps u_{i-1} vectors of the basis to the corresponding vectors in C_{i-1} and is trivial on the remaining vectors of the basis. Thus $\mathrm{Ker}\left(\partial : C_i \to C_{i-1}\right)$ is generated by the last $u_i + b_i$ vectors of the basis but only b_i of them give a contribution to the homology while u_i remaining ones generate the image of $\partial : C_{i+1} \to C_i$

In our case, the alternated sum of Betti numbers provides the equality

$$\chi = 2 + b_2 \,,$$

while from the Morse complex (formula (11) on p. 481) we get

$$\chi = n^3 - 4n^2 + 6n \,.$$

Therefore, $b_2 = n^3 - 4n^2 + 6n - 2$:

n	1	2	3	4	5	6	7
b_2	1	2	7	26	53	106	187

.

We have used Main Lemma and Corollary 1, which hold only for $n > 2$. The proof for the exceptional cases, $n = 1$ and $n = 2$, is given below. $\square$

In the case $n = 1$, the surface $M^2 = \mathbb{C}\mathrm{P}^2$ has Betti numbers $(1, 0, 1, 0, 1)$. Hence, $b_2 = 1$ and $\chi = 3$, confirming the above general formula.

In the case $n = 2$, the surface M^2 is "the complexified sphere" in $\mathbb{C}\mathrm{P}^3$ given by the equation $x^2 + y^2 + z^2 = 1$.

Theorem. *The complexified sphere is diffeomorphic, as a real 4-manifold, to the product $\mathbb{S}^2 \times \mathbb{S}^2$ of two real 2-spheres.*

Consequently, the Betti numbers are $(1, 0, 2, 0, 1)$, confirming again the above general formula. It is easy to construct on M^2 a smooth function with one maximum, one minimum and two critical points of index 2.

Remark. The real version of the above theorem is the well-known fact that a one-sheet hyperboloid of $\mathbb{R}\mathrm{P}^3$, for instance, that given by the equation

$$x^2 + y^2 = 1 + z^2,$$

is diffeomorphic to the torus $\mathbb{S}^1 \times \mathbb{S}^1$ (Fig. 12.7).

The statement of the above theorem is explained by the fact that the Riemann sphere $\mathbb{C}\mathrm{P}^1 = \mathbb{S}^2$ is the complex version of the circle $\mathbb{R}\mathrm{P}^1 = \mathbb{S}^1$.

Proof of the Theorem. Write the equation in the form $xy = zw$, for suitable homogeneous coordinates. Rewrite it as the proportion

$$\frac{x}{z} = \frac{w}{y} \,.$$

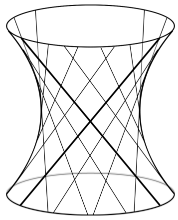


Figure 12.7: The torical structure of the hyperboloid is defined by its two fibrations into projective lines.

Denoting by a this ratio, we get

$$x = az \ , \quad w = ay \qquad (a \in \mathbb{C}\mathrm{P}^1) \ .$$

We have thus defined a projection of M^2 to $\mathbb{C}\mathrm{P}^1 = \mathbb{S}^2$.

Rewriting the equation in the form of a different proportion,

$$\frac{x}{w} = \frac{z}{y} = b \ ,$$

we get

$$x = bw \ , \quad z = by \qquad (b \in \mathbb{C}\mathrm{P}^1) \ .$$

This defines a second projection to $\mathbb{S}^2$.

For each a and b, taking $x = 1$, we reconstruct $z = 1/a$, $w = 1/b$, $y = 1/ab$. So, our two maps provide a diffeomorphism $M^2 \approx \mathbb{C}\mathrm{P}^1 \times \mathbb{C}\mathrm{P}^1$, which, moreover, is holomorphic: The complex surface M^2 is holomorphically diffeomorphic to the product of the two Riemannian spheres. □

12.6 Betti numbers of hypersurfaces in $\mathbb{C}\mathrm{P}^m$

In order to compute the Betti numbers of a smooth algebraic hypersurface M of degree n in $\mathbb{C}\mathrm{P}^m$, we are going to use the long exact sequence of the pair $(\mathbb{C}\mathrm{P}^m, M)$ - see p. 425. It relates the required homology $H_k(M)$, the known homology $H_k(\mathbb{C}\mathrm{P}^m)$ and the relative homology $H_k(\mathbb{C}\mathrm{P}^m, M)$:

$$\to H_{k+1}(\mathbb{C}\mathrm{P}^m, M) \xrightarrow{D} H_k(M) \xrightarrow{i_*} H_k(\mathbb{C}\mathrm{P}^m) \xrightarrow{p_*} H_k(\mathbb{C}\mathrm{P}^m, M) \to \dots \ . \qquad (13)$$

By Poincaré duality, $H_k(\mathbb{C}\mathrm{P}^m, M)$ is isomorphic to the absolute cohomology $H^{2m-k}(\mathbb{C}\mathrm{P}^m \smallsetminus M)$ of the complement $\mathbb{C}\mathrm{P}^m \smallsetminus M$. Moreover, we have the

Lemma (proved below)**.** *The group* $H^{2m-k}(\mathbb{C}\mathrm{P}^m \setminus M)$ *is trivial for* $k < m$.

The general "Stein property" of an open complex m-dimensional manifold $\mathbb{C}\mathrm{P}^m \setminus M$, mentioned on p. 474, is this vanishing of the cohomologies of degree $K > m$ $(K = 2m - k,\ k < m)$. This enables us to prove the following

Theorem 2. *The Betti numbers of a complex hypersurface* M *of degree* n *in* $\mathbb{C}\mathrm{P}^m$ *are*

$$b_{2k} = 1 \quad \textit{and} \quad b_{2k-1} = 0 \quad (\text{except for } b_{m-1})\,,$$

$$b_{m-1} = \begin{cases} m - \chi & \textit{if } m \textit{ is even} \\ \chi - (m-1) & \textit{if } m \textit{ is odd}\,, \end{cases}$$

where $\chi = m + \frac{n-1}{n} + \frac{(1-n)^{m+1}}{n}$.

Proof. The lemma and the exact sequence (13) provide the short sequences

$$0 \to H_k(M) \to H_k(\mathbb{C}\mathrm{P}^m) \to 0$$

for $k < m - 1$. Therefore the Betti numbers b_i of M repeat those of $\mathbb{C}\mathrm{P}^m$ for $i < m - 1$ (it is the so-called "Lefschetz hyperplane theorem"). Thus we obtain the first half of the Betti numbers.

The other half $(i > m - 1)$ is provided by the Poincaré duality, claiming that $b_i = b_{2(m-1)-i}$.

Therefore we have $b_{2k} = 1$ and $b_{2k-1} = 0$ except for the middle Betti number b_{m-1}, which is computable from the Euler characteristic χ obtained in Theorem 1 (p. 481). Namely, since $\chi = \sum(-1)^i b_i$ we have

$$\chi = m - b_{m-1} \text{ if } m \text{ is even}\,; \qquad \chi = (m-1) + b_{m-1} \text{ if } m \text{ is odd.} \qquad \square$$

Remark (Newton polyhedra). In the case where m is even, the Betti number $b_{m-1}(M)$, calculated by the Newton polyhedron theory, is equal to the number of integral points strictly inside the cube $[0, n]^m$, which do not belong to the hyperplanes that contain its vertices and are orthogonal to the main diagonal of the cube (that is, the hyperplanes $k_1 + \cdots + k_m = jn$, $0 \le j \le m$).

Example. For $m = 2$ we count the integral points strictly inside the square, which do not belong to its diagonal. Their number is $(n-1)^2 - (n-1) = (n-1)(n-2)$. It coincides with the Betti number $b_1 = 2g$ of the plane algebraic curve of degree n (see p. 148-149).

Exercise. For $m = 4$ we have to count the integral points outside 3 hyperplanes strictly inside the 4-cube $[0, n]^4$. Check that the formula of Theorem 2 coincides with this number of points

$$\begin{aligned} b_3 &= (n-1)^4 - 2\binom{n-1}{3} - \frac{1}{3}\,(2n^3 - 6n^2 + 7n - 3) \\ &= n^4 - 5n^3 + 10n^2 - 10n + 4. \end{aligned}$$

Proof of the lemma. Consider the *Veronese embedding* $\mathbb{C}\mathrm{P}^m \to \mathbb{C}\mathrm{P}^N$ (with $N = \binom{n+m}{m} - 1$) which in homogeneous coordinates is given by all possible monomials of degree n. Under this embedding, the degree n hypersurface M turns into a hyperplane section of the embedded image of $\mathbb{C}\mathrm{P}^m$. Hence, the complement $\mathbb{C}\mathrm{P}^m \setminus M$ admits a closed embedding to the affine space $\mathbb{C}^N$. Its topology can be studied using the Morse function r^2. This function has no critical point of Morse index $> m$. In consequence, the Morse cochain complex has no generators of degree $> m$, and hence the corresponding cohomologies are trivial. □

12.6.1 The real part of a complex quadratic form

Problem. Describe the real quadratic forms Φ in $\mathbb{R}^{2\ell} = \mathbb{C}^\ell$, which are representable as real parts of complex quadratic forms:

$$\Phi = \mathrm{Re}\, Q\ , \quad Q : \mathbb{C}^\ell \to \mathbb{C} \text{ being a complex quadratic form.}$$

Solution. The dimension of the real vector-space of the forms $\mathrm{Re}\, Q$ is equal to the real dimension of the space of complex quadratic forms Q: $\ell(\ell+1)$.

Indeed, if $\mathrm{Re}\, Q = 0$, the form Q is everywhere 0; otherwise, if Q would take a non-zero value v at some vector $z \in \mathbb{C}^\ell$, then the complex form Q would take the (non-zero) real value 1 at the vector $z/\sqrt{v} \in \mathbb{C}^\ell$.

Theorem. *A real quadratic form Φ on $\mathbb{R}^{2\ell} = \mathbb{C}^\ell$ is representable as real part of a complex quadratic form on $\mathbb{C}^\ell$ if and only if $\Phi(iw) = -\Phi(w)$.*

The condition $\Phi(iw) = -\Phi(w)$ being evidently necessary for the representability $\Phi = \mathrm{Re}\, Q$ we shall prove below that it is also sufficient.

Indeed, for the real vectors $p \in \mathbb{R}^\ell$, $q \in \mathbb{R}^\ell$, forming $w = p + iq \in \mathbb{C}^\ell$, the above condition on Φ takes the form $\Phi(-q+ip) = -\Phi(p+iq)$. Let $\left(\begin{smallmatrix} A & B \\ B' & C \end{smallmatrix}\right)$ be the $2\ell \times 2\ell$ real matrix representing the real form $\Phi : \mathbb{R}^{2\ell} \to \mathbb{R}$,

Lemma. *The conditions $C = -A$ and $B = B'$ hold.*

Proof. The identity $\Phi(-q+ip) = -\Phi(p+iq)$ together with the equalities

$$\Phi(p+iq) = (Ap,p) + (Bp,q) + (B'q,p) + (Cq,q)\,,$$

$$\Phi(-q+ip) = (Aq,q) - (Bq,p) - (B'p,q) + (Cp,p)\,,$$

provide the identities

$$(Ap, p) = -(Cp, p) \quad \text{and} \quad (Bq, p) + (B'p, q) = (Bp, q) + (B'q, p)\,,$$

which mean that $2B = 2B'$ and $A + C = 0$. $\square$

The real dimension of the space of pairs of real symmetric matrices A and B of order ℓ is equal to $2\,\frac{\ell(\ell+1)}{2} = \ell(\ell+1)$. Therefore, all such pairs of matrices provide (bijectively) all real forms that are representable as the real part of a complex quadratic form $\Phi = \operatorname{Re} Q$.

Proof of the Theorem. A real quadratic form $\Phi : \mathbb{R}^{2\ell} \to \mathbb{R}$ satisfying the condition $\Phi(iw) = -\Phi(w)$ can be written by the normal diagonal expression

$$\Phi = \sum_{j=1}^{\ell} c_j(p_j^2 - q_j^2) \quad (\text{all } c_j \geqslant 0)$$

in some Euclidean orthonormal coordinate system.

Choose the first ℓ Hermitian orthonormal basic vectors that correspond to the p_j coordinates, as the Hermitian basis of the complex vector-space $\mathbb{C}^\ell$.

In the corresponding complex coordinate system $\{w_j\}$, the real quadratic form Φ has the expression

$$\Phi = \operatorname{Re} \sum_{j=1}^{\ell} c_j w_j^2\ .$$

Consequently, for these coordinates $Q = \sum_{j=1}^{\ell} c_j w_j^2$. $\square$

The above description of the structure of the form Φ in terms of a pair of real symmetric matrices is equivalent to the following proposition.

Proposition. *Any complex quadratic form in the Hermitian complex vector-space $\mathbb{C}^\ell$ is unitary equivalent to the diagonal form $Q = \sum_{j=1}^{\ell} c_j z_j^2$ with real constants $c_j \geqslant 0$.*

Observe that the real dimension $\ell^2 + \ell$ of the space of the complex quadratic forms in $\mathbb{C}^\ell$ equals the real dimension ℓ^2 of the unitary group $\mathrm{U}(\ell)$ plus the number ℓ of invariants c_j in the normal form.

Remark. The above normal forms of the quadratic part of the Taylor series of a real smooth function on a complex ℓ-manifold were extended to higher degree terms by Chern and Moser [58, 59]. It would be nice to deduce from their normal forms the description of the generic bifurcations of the Morse complex of the function r^2 restricted to holomorphic submanifolds (of the Hermitian space $\mathbb{C}^m$) depending generically on the parameters.

Chapter 13

Topologic Abel's theorem on unsolvability of algebraic equations

One of the first and most important impossibility results in mathematics is Abel's Theorem: *There exists no finite combination of radicals and rational functions solving the generic algebraic equation of degree* 5 (*or higher than* 5).

The topological proof of Abel's theorem presented below was given first time by Arnold in 1964 during a semester of High School lectures at Moscow. Let us describe it.

Our study of regular polyhedra symmetry groups, in Chapter 2, led us to the five Kepler cubes inscribed into the dodecahedron (Fig. 2.14, p.50). We use these cubes to obtain the natural isomorphism between the dodecahedron rotation group and the group of the 60 even permutations of five elements (the five Kepler cubes).

The fact that there is no non-trivial normal subgroup in the rotation group of the dodecahedron (an easy result of elementary geometry) implies the non-solvability (defined below) of the group of permutations of five elements.

This non-solvability, combined with the topological study of the monodromies of the ramified coverings, provides immediately the topological proof of Abel's Theorem. Namely, the monodromy group of any finite combination of radicals is solvable (because the radical monodromy is a cyclic commutative group) whilst the monodromy group of the algebraic function $z(a)$ defined by the quintic equation $z^5 + az + 1 = 0$ is the non-solvable group formed by the 120 permutations of the 5 roots of the equation.
(Notice that all permutation groups of less than five elements are solvable – this solvability being responsible for the solvability of the equations of degree smaller than 5.)

The topological unsolvability argument provides more than Abel's Theorem. Namely, no function having the same topological branching type as $z(a)$ is representable as a finite combination of rational functions and radicals.

13.1 Solving algebraic equations by radicals

Multivalued functions. One lears in school that $\sqrt{4}$ has the two values $\{-2, 2\}$. Similarly, for any complex number $a \neq 0$ the function $z = \sqrt{a}$ has two values, say $\{-z_a, z_a\}$, taking the single value $z = 0$ (of multiplicity 2) only for $a = 0$. Along this chapter we shall consider multivalued functions.

Everyone knows the formula providing the two roots of a quadratic equation in terms of the square root and rational operations on the coefficients. For the equation $z^2 + 2az + b = 0$ it provides the two-valued function:

$$z(a, b) = -a \pm \sqrt{a^2 - b}\,. \tag{1}$$

Similar formulas for the equations of degrees three and four were discovered in the 16th Century by italian mathematicians. For example, to solve the cubic equation $y^3 + ay^2 + by + c = 0$ by radicals, one starts by putting $y = z - a/3$ to transform it to

$$z^3 + 3pz + 2q = 0\,, \tag{2}$$

whare p and q are polynomials of a, b and c. The formula solving this equation, found by Niccolo Fontana Tartaglia, is the 3-valued complex function

$$z(p, q) = \sqrt[3]{-q + \sqrt{q^2 + p^3}} + \sqrt[3]{-q - \sqrt{q^2 + p^3}}\,, \tag{3}$$

where for each value of the first cubic root the value of the second one is chosen so that their product be equal to p.

Abel was trying several years to find a formula to solve equations of degree five. The result of this long work was not a formula but the remarkable

Abel Theorem. *Equations of degree* 5 *cannot be solved in terms of any radicals and rational functions of the coefficients.*

Abel's statement is true already for the special equation

$$z^5 + az + 1 = 0$$

whose five complex solutions form a multivalued algebraic function $z(a)$. We shall prove that *this function is not representable in the form of a (finite) combination of radicals and rational functions by topological reasons*: Its *monodromy group* (which describes the permutations of the 5 solutions when the parameter a varies) is topologically different from the monodromy of any finite combination of radicals and of rational operations.

13.2 Monodromy Group

The *monodromy group* of an n-valued function is the representation of the fundamental group of the complement to the discriminant (in which at least two values coalesce) into the group of permutations of the n values of the function. Hence it is a subgroup of the group $S(n)$ of permutations of n elements (the n values of the function).

To understand the monodromy of the algebraic functions, we shall start with simple examples, depicting their Riemann surfaces (see below).

PROBLEM. Find the monodromy group of the algebraic function $z = \sqrt[n]{a}$.

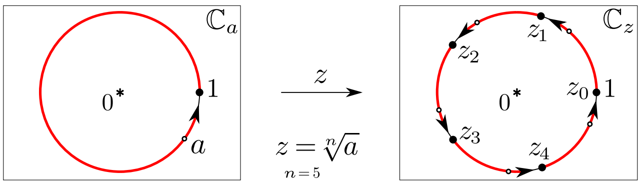
SOLUTION. This n-valued complex function z can be prolonged continuously from any value if the path followed by the argument a does not pass through the point $a = 0$ in $\mathbb{C}$. In other words, the function defines an n-fold covering of the domain $\mathbb{C} \setminus 0$ of the variable a, which is a locally trivial fibration whose fibre consist of n points.

We shall start at the point $a = 1$. Consider the n roots of degree n of 1:

$$z_k = e^{k\frac{2\pi}{n}i}, \quad k = 0, \ldots, n-1.$$

If a turns along the unit circle, $a(t) = e^{2\pi i t}$, $0 \leqslant t \leqslant 1$, then the root z_k moves along the arc $\zeta_k(t) = e^{(k+t)\frac{2\pi}{n}i}$:

$a(t) = e^{2\pi i t}$



$\zeta_k(t) = e^{(k+t)\frac{2\pi}{n}i}$.

Clearly $\zeta_k(0) = z_k$ and $\zeta_k(1) = z_{k+1}$ (understanding z_n as z_0).

In other words, the monodromy of the function $\sqrt[n]{a}$ is the cyclic permutation of the n roots of degree n of $a = 1$.

The path $e^{2\pi i t}$, $0 \leqslant t \leqslant 1$, is a generator of the fundamental group of the complement to the ramification manifold, which consists of the point $a = 0$:

$$\pi_1(\mathbb{C} \setminus 0, 1) = \mathbb{Z}.$$

The whole monodromy group, which represents the fundamental group of the complement to the ramification manifold by permutating the elements

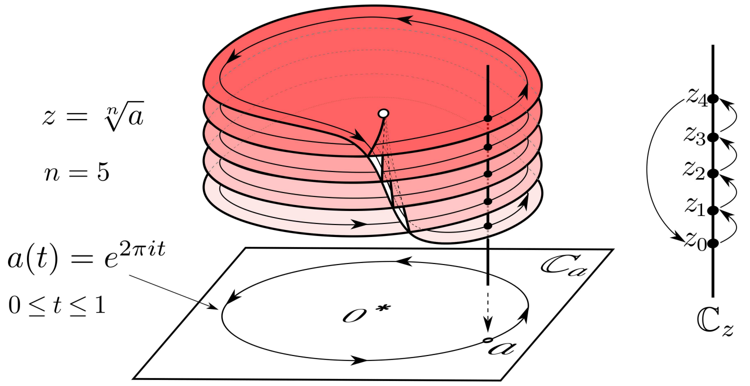


Figure 13.1: Monodromy and Riemann surface of the function $z(a) = \sqrt[n]{a}$.

of the fibre $\mathbb{Z}_n = \{\zeta_k(0) = z_k\}$, is therefore the cyclic group of permutations of the n roots $\zeta_k(0) = z_k$, generated by the rotation (Fig. 13.1) :

$$M : \mathbb{Z}_n \to \mathbb{Z}_n, \quad Mz_k = z_{k+1}.$$

This cyclic group G consists of n elements $G = \{1, M, M^2, \dots, M^{n-1}\}$, $M^n = 1$. We have thus proved the

Theorem 1. *The monodromy group of the algebraic function $z = \sqrt[n]{a}$ is the cyclic commutative group $\mathbb{Z}_n$ of the rotations of a regular n-gon,*

$$\{z_k = e^{k\frac{2\pi i}{n}}, \quad 0 \leqslant k < n\}.$$

Riemann surfaces. In this chapter, by *Riemann surface* of a multivalued complex funcion $z = z(a)$ we mean its *multigraph*, which is the subset of $\mathbb{C} \times \mathbb{C}$ formed by the "pairs" $(a, z(a))$ where $z(a)$ is the subset of $\mathbb{C}$ formed by all values of the function z at a. So the Riemann surface is the branched covering over $\mathbb{C} = \{a\}$ whose sheets over a point a_0 represent the complex values of the multivalued function $z = z(a)$ at $a = a_0$.

Representing this surface (of $\mathbb{C} \times \mathbb{C} \approx \mathbb{R}^4$) in $\mathbb{R}^3$, its sheets cannot be joined without self-intersection along some curves (as we cannot avoid such self-intersections to depict the Klein bottle or the projective space in $\mathbb{R}^3$), but such self-intersection curves are only apparent. In these pictures, the vertical line over a point a represents the complex line $\mathbb{C} = \{z\}$, where the depicted points $\{z_0, \dots, z_{n-1}\}$ are the values of the function $z = z(a)$ and the curved arrows indicate the permutation provided by the given loop. See Fig. 13.1.

Problem. Find the monodromy group of the bi-valued function defined by the roots of the equation $z^2 - az + 1 = 0$, that is, $z(a) = \frac{a}{2} + \sqrt{(\frac{a}{2})^2 - 1}$.

Answer. The monodromy group is the whole group $S(2) = \mathbb{Z}_2$.

Solution. The discriminant being $(a/2)^2 - 1$, for each value $a \neq \pm 2$, the function z takes two different values, say, $z(a) = \{u, v\}$ with $u \neq v$.

(λ) To start we move the parameter a along a path λ from 0 to $2+\varepsilon$ (with $\varepsilon > 0$) on the positive real line – Fig. 13.2 λ. When $a \leqslant 2$ the roots u, v describe two complex-conjugate arcs on the unit circle (starting at the initial roots $z(0) = \{-i, i\}$) because $z(a) = a/2 \pm i\sqrt{1-(a/2)^2}$. At $a = 2$ these two roots converge into the root of multiplicity two $z = 1$, which splits into two inverse real numbers when $a > 2$ (because $uv = 1$).

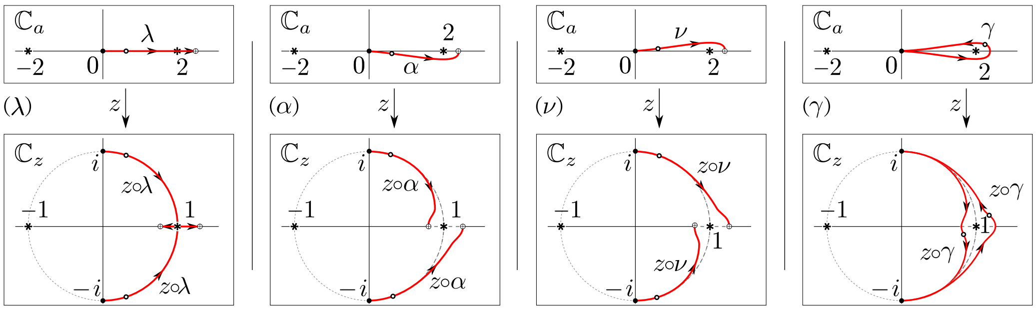


Figure 13.2: The loop γ around the branching point $a = 2$ permutes the values $z(0) = \{i, -i\}$ of the bi-valued complex function $z(a) = a/2 + \sqrt[2]{(a/2)^2 - 1}$.

(α) Now we slightly deform the real path $a = \lambda(t)$ into a path $a = \alpha(t)$ having negative imaginary part except at its end points, $\alpha(0) = 0$ and $\alpha(2+\varepsilon) = 2+\varepsilon$ – Fig. 13.2 α. In this case, one of the curves of roots lies inside the unit circle, while the other lies outside, because $uv = 1$, which means that one of the curves of roots is obtained from the other by complex conjugation followed by an inversion with respect to the unit circle.

(ν) The curve $a = \nu(t)$ is obtained deforming λ but taking positive imaginary part. Its curves of roots behave like those of α (see Fig. 13.2 ν).

(γ) We form the loop $a = \gamma(t)$ which tourns around 2 by going along α from 0 to $2+\varepsilon$ and then going back to 0 along ν but taking its opposite orientation. One of the curves of roots, say $z(a) = u$ starting at $-i$ and ending at i, lies outside the unit circle, while the other one, $z(a) = v$ starting at i and ending at $-i$, lies inside. Therefore the loop γ permutes the roots $-i$ and i (see Fig. 13.2 γ).

Multiplying the loop γ by -1 we get a new loop $-\gamma$ around $a = -2$ which also permutes $-i$ and i. The action of the loops γ and $-\gamma$ on the roots over the initial point $a = 0$, $z_0 = -i$ and $z_1 = i$, is depicted on Fig. 13.3.

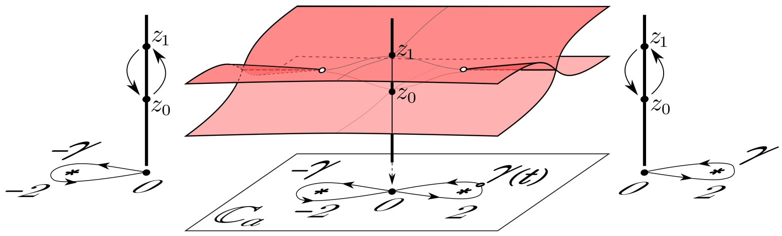


Figure 13.3: Monodromy and Riemann surface of $z(a) = a/2 + \sqrt[2]{(a/2)^2 - 1}$.

Problem. Find the monodromy group of the three-valued algebraic function $z(a)$ defined by the complex roots of the equation $z^3 - az - 2 = 0$.

Hint. 1. Since the discriminant is provided by the formula which solves this equation,

$$z(a) = \sqrt[3]{1 + \sqrt[2]{1 - (a/3)^3}} + \sqrt[3]{1 - \sqrt[2]{1 - (a/3)^3}},$$

it consists of the three cubic roots of 27: $\{3, r_1 = 3e^{i2\pi/3}, r_2 = 3e^{i4\pi/3}\}$.
2. The values of $z(0)$ are the cubic roots of 2: $z_0 = \sqrt[3]{2}$, $z_1 = \sqrt[3]{2}e^{i2\pi/3}$, $z_2 = \sqrt[3]{2}e^{-i2\pi/3}$.
3. Moving continuously the parameter a from 0 to 3, the two roots z_1, z_2 will converge to the double real root $z = -1$, because our equation becomes $(z+1)^2(z-2) = 0$ (verify it!).
4. A loop μ around $a = 3$ (like γ above) permutes the roots z_1 and z_2 – Fig. 13.4.

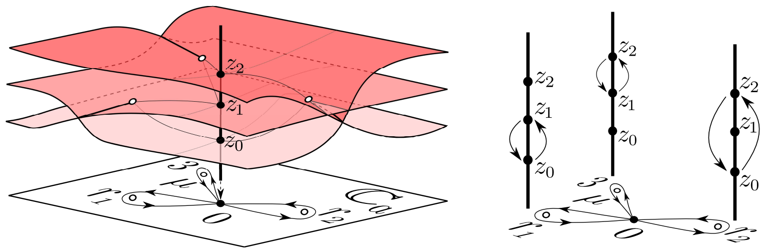


Figure 13.4: The monodromy and Riemann surface of the three-valued function $z(a) = \sqrt[3]{1 + \sqrt{1 - (a/3)^3}} + \sqrt[3]{1 - \sqrt{1 - (a/3)^3}}$.

5. Multiplying μ by $e^{i2\pi/3}$ and by $e^{-i2\pi/3}$, we get one loop around $a = r_1$ and other around $a = r_2$, which respectively produce the permutations $z_2 \leftrightarrow z_0$ and $z_0 \leftrightarrow z_1$.

Example. Consider the algebraic function $z(a)$ defined by

$$z(a) = \sqrt{2 - \sqrt[n]{a}}\,. \tag{4}$$

This algebraic function has (generically) $2n$ different values for a fixed (complex) value of the argument a. It defines a ramified covering of the domain $\mathbb{C}$ of the variable a. The ramification points occur at the values of a for which one has to compute some root of 0 when applying formula (4).

The ramification values of the argument a are therefore $a = 0$ and the points a for which one of the values of $\sqrt[n]{a}$ equals 2. The last condition means that $a = 2^n$. Thus, our function defines a $2n$-fold covering of the plane $\{a\} = \mathbb{C}$, ramified at the two points 0 and 2^n.

The fundamental group of the base manifold $B = \mathbb{C} \setminus \{0, 2^n\}$ of the covering is the free group with two generators: one is the loop $\{a_1(t)\}$ making one simple turn around the ramification point $a = 0$, the other, $\{a_2(t)\}$, makes a turn around the ramification point 2^n.

We can decompose the $2n$ values of the function $z(a)$ at the initial base point $a = 1 \in B$ into n numbered sets of two elements (pairs): The pair number k, $\{+\sqrt{2 - z_k}\,, -\sqrt{2 - z_k}\,\}$, corresponds to the value $z_k = e^{2\pi k/n}$ of $\sqrt[n]{a = 1}$. Denote by $\{u_k, v_k\}$ this unordered pair of values of the function (4).

For every continuous path of a in B, the continuation of the pair $\{u_k, v_k\}$ is well defined, since z_k is continued unambiguously as $\sqrt[n]{a}$. Therefore, returning to $a = 1$ we shall arrive to a permutation of the $2n$ values of the function (4) that has the following structure: The pairs $\{u_k, v_k\}$ are permuted cyclically according to the monodromy group of the function $\sqrt[n]{a}$.

So, the monodromy of the function (4) produces a cyclic permutation of pairs that sends each pair $\{u_k, v_k\}$ to another pair $\{u_\ell, v_\ell\}$. However, u_k can be sent to u_ℓ or to v_ℓ – in the first case, v_k is sent to v_ℓ and in the second, to u_ℓ. Different cases may occur for distinct loops in the base space B.

In other words, the monodromy group G_1 of the function (4) is sent homomorphically onto the monodromy group $G_0 = \mathbb{Z}_n$ of the function $\sqrt[n]{a}$. The elements of the kernel of this homomorphism (being also the monodromy action of those elements of the fundamental group $\pi_1(\mathbb{C} \setminus 0)$) fix all pairs, but may permute the two elements of each pair $\{u_k, v_k\}$ in some way, which depends on the loop whose monodromy representation we are studying.

We get therefore the exact sequence

$$1 \to \mathrm{Ker} \to G_1 \to G_0 \to 1\,, \tag{5}$$

where 1 means the trivial group consisting of one element, the group Ker is a subgroup of $(\mathbb{Z}_2)^n$ and the group G_0 is $\mathbb{Z}_n$.

In our particular example, it is not so difficult to compute the groups Ker and G_1 completely, but we shall not do it, since for the general Abel theorem we shall need only a small generalisation of the exact sequence (5).

The main idea is that the representation of the algebraic function (4) in terms of roots of degrees n and 2 is reflected at the monodromy level: The monodromy group of the algebraic function (4) is some "combination" of the commutative groups $G_0 = \mathbb{Z}_n$ and $\mathrm{Ker} \subset \mathbb{Z}_2^n$. An exact sequence of type (5) may describe the direct product G_1 of the groups G_0 and Ker, but it may also happen for some non-commutative "combinations" of the commutative groups G_0 and Ker.

To describe this "combination" in the general situation, we have to introduce the notion of "solvable" and "unsolvable" groups.

13.3 Solvable and unsolvable groups

Definition. A group G is said to be *solvable*, if it has a chain of subgroups

$$(H_N = e) \subset H_{N-1} \subset \ldots \subset H_1 \subset (G = H_0),$$

with the following two properties:

1. H_j is an invariant subgroup in H_{j-1};
2. the quotient $C_j = H_{j-1}/H_j$ is commutative, for $j = 1, 2, \ldots, N$.

In a sense, the solvable group G is a "combination" of the commutative groups C_j. But the group G itself may be non-commutative.

Example. The group $S(3)$ of the 6 permutations of 3 elements (or of the symmetries of an equilateral triangle) is solvable and non-commutative.

Proof. Consider the chain

$$(H_2 = e) \subset \mathbb{Z}_3 \subset S(3),$$

where $\mathbb{Z}_3$ means the subgroup of 3 rotations in the symmetry group of the equilateral triangle, that is, the subgroup $S^+(3)$ of even permutations in $S(3)$.

Since the groups $C_1 = \mathbb{Z}_2$ and $C_2 = \mathbb{Z}_3$ are commutative, the group $S(3)$ is solvable. □

Problem. Prove that the group $S(4)$ of the 24 permutations of 4 elements (or of the symmetries of a tetrahedron) is solvable.

Solution. The subgroup $S^+(4)$ of even permutations contains 12 elements, which represent the tetrahedron rotations. It is an invariant subgroup* which is the kernel of the natural map $S(4) \to \mathbb{Z}_2$, representing the action of a symmetry on the two orientations of the tetrahedron.

Among the 12 rotations of the tetrahedron, there are four distinguished rotations which form an invariant subgroup. To construct this subgroup, join the middle points of each pair of opposite edges of the tetrahedron by a "joining line". The rotations permute the three joining lines, defining a homomorphism $S^+(4) \to S(3)$ whose kernel, as every homomorphism kernel, is an invariant subgroup. It consists of four rotations : The identity map and the rotation of $180°$ around each of the joining lines.

Each of these three non-trivial rotations is an involution (its square being the identity map) and the product of any two of them equals the third. Thus, counting the identity, the four distinguished rotations form the commutative group $\mathbb{Z}_2^2$.

We get the chain of subgroups

$$e \hookrightarrow \mathbb{Z}_2^2 \hookrightarrow S^+(4) \hookrightarrow S(4),$$

(consisting of 1, 4, 12 and 24 elements, respectively) which verify both conditions of the definition of solvability: The commutative quotient groups are

$$C_1 = \mathbb{Z}_2, \quad C_2 = \mathbb{Z}_3, \quad C_3 = \mathbb{Z}_2^2.$$

This proves that the groups $S^+(4)$ and $S(4)$ are solvable.

Problem. Is the group $B(3)$ of the 48 symmetries of the cube solvable?

Solution. The 24 rotations form a solvable invariant subgroup, and we get the chain

$$e \hookrightarrow \mathbb{Z}_2^2 \hookrightarrow S^+(4) \hookrightarrow S(4) \hookrightarrow B(3),$$

proving the solvability.

Problem. Can a subgroup of a solvable group be non-solvable?

Answer. *Any subgroup of a solvable group is solvable.*

*As is invariant any subgroup containing exactly one half of the elements of the group.

Solution. The intersections of a given subgroup $K \subset G$ with the subgroups H_j form a chain of subgroups $(K_j = K \cap H_j) \subset (K_{j-1} = K \cap H_{j-1})$.

The intersection K_j is an invariant subgroup of the intersection K_{j-1}. Indeed, for $a \in K_j$, $b \in K_{j-1}$ we have $bab^{-1} \in H_j$, the subgroup H_j is invariant in H_{j-1} and also $bab^{-1} \in K$ (since $a \in K$, $b \in K$). Therefore K_j is an invariant subgroup of K_{j-1}.

Problem. Can a homomorphism send a solvable group onto a non-solvable group?

Solution. The images $\widetilde{H}_j = f(H_j)$ of the subgroups H_j under the homomorphism $f : G \to \widetilde{G}$ onto $\widetilde{G}$ form a chain

$$(\widetilde{H}_N = e) \hookrightarrow \widetilde{H}_{N-1} \hookrightarrow \ldots \hookrightarrow \widetilde{H}_1 \hookrightarrow (\widetilde{H}_0 = \widetilde{G}).$$

The subgroup $\widetilde{H}_j$ of $\widetilde{H}_{j-1}$ is invariant: If $\mathtt{a} \in \widetilde{H}_j$, $\mathtt{b} \in \widetilde{H}_{j-1}$, we can represent them as $\mathtt{a} = f(a)$ with $a \in H_j$ and $\mathtt{b} = f(b)$ with $b \in H_{j-1}$; then $\mathtt{bab}^{-1} = f(b)f(a)f(b^{-1}) = f(bab^{-1})$ and $bab^{-1} \in H_j$ because H_j is an invariant subgroup of H_{j-1}. Therefore $f(bab^{-1}) \in f(H_j) = \widetilde{H}_j$ and thus $\mathtt{bab}^{-1} \in \widetilde{H}_j$ for any $\mathtt{b} \in \widetilde{H}_{j-1}$, $\mathtt{a} \in \widetilde{H}_j$, proving the invariance of $\widetilde{H}_j$ in $\widetilde{H}_{j-1}$.

Since $\widetilde{H}_j = f(H_j)$, we have that the j-th quotient group $\widetilde{C}_j = \widetilde{H}_{j-1}/\widetilde{H}_j$ is the image of the quotient $C_j = H_{j-1}/H_j$, under the homomorphism $C_j \to \widetilde{C}_j$ induced by the homomorphism f.

Hence, the quotient groups $\widetilde{C}_j$ are commutative because they are homomorphic images of commutative groups.

Therefore the image of a solvable group under any homomorphism of groups is also a solvable group.

Problem. Can a homomorphism send a solvable group *into* a non-solvable group?

Answer. Yes. Indeed, the trivial homomorphism onto the element e of any group defines a homomorphism *into* that group.

Arnold's geometrical lectures of 1964 were published in 1976 in [1] by one of the pupils of High School audience, V.B. Alekseev, who has somewhere algebraised them and added the wrong statement that the homomorphisms into non-solvable groups exist only from non-solvable groups (p. 103 of [1]). He added also that "all algebraic functions are (everywhere) analytic" (p. 101 of [1]). It is unknown whether $\sqrt{z}$ is an analytic or a non-algebraic function for him. He needs this analyticity for some (strange) proofs of true non-trivial facts (missing in Arnold's lectures, read to high-school students unaware of analysis).

Problem. Prove that a group G is solvable if there exists a homomorphism from G onto a solvable group B, $f : G \to B$, whose kernel K is solvable (providing an exact sequence $1 \to K \to G \xrightarrow{f} B \to 1$).

Solution. Start from the chain of subgroups

$$(e = H_N) \hookrightarrow H_{N-1} \hookrightarrow \ldots \hookrightarrow H_1 \hookrightarrow (H_0 = B),$$

given by the solvability of the group B, whose preimages, $f^{-1}H_j = \widehat{H}_j$, form the chain

$$K = \widehat{H}_N \hookrightarrow \widehat{H}_{N-1} \hookrightarrow \ldots \hookrightarrow \widehat{H}_1 \hookrightarrow (\widehat{H}_0 = G)\,.$$

This new chain may be continued from the left hand side by the chain of subgroups of the solvable group K.

We shall prove that the resulting chain of subgroups of the group G verify both conditions of the definition of solvability.

First, to prove that the subgroup $\widehat{H}_j$ is invariant in $\widehat{H}_{j-1}$ we consider $a \in \widehat{H}_j$ and $b \in \widehat{H}_{j-1}$, observing that $f(bab^{-1}) = f(b)f(a)f(b^{-1}) \in H_j$ because $f(a) \in H_j$ and $f(b) \in H_{j-1}$ (by the definitions of $\widehat{H}_j$ and $\widehat{H}_{j-1}$). Thus for any $a \in \widehat{H}_j$ and $b \in \widehat{H}_{j-1}$ we have that

$$bab^{-1} \in f^{-1}(f(bab^{-1})) \in f^{-1}H_j = \widehat{H}_j,$$

proving the invariance of $\widehat{H}_j$ in $\widehat{H}_{j-1}$.

Next, we recall that the quotient group G/H of any group G by its invariant subgroup H is commutative if and only if for any pair of elements a, b of G the product $aba^{-1}b^{-1}$ belongs to H.

Hence to prove the commutativity of the quotient group $\widehat{C}_j = \widehat{H}_{j-1}/\widehat{H}_j$, we shall take the products $aba^{-1}b^{-1}$ for a and b in $\widehat{H}_{j-1}$ to obtain that $f(aba^{-1}b^{-1}) = \mathtt{aba}^{-1}\mathtt{b}^{-1}$, where $\mathtt{a} = f(a)$ and $\mathtt{b} = f(b)$ belong to H_{j-1}. Therefore $\mathtt{aba}^{-1}\mathtt{b}^{-1} \in H_j$, by the commutativity of $C_j = H_{j-1}/H_j$.

Returning to the group G, we see that

$$aba^{-1}b^{-1} \in f^{-1}(H_j) = \widehat{H}_j,$$

whatever be the elements a, b in $\widehat{H}_{j-1}$, implying the commutativity of the quotient group $\widehat{C}_j = \widehat{H}_{j-1}/\widehat{H}_j$.

Thus, using the chains of subgroups of B and K we have constructed a longer chain of subgroups of the group G for which the quotients are commutative. Therefore *the extension G of the subgroup K is a solvable group.*

Proposition. *The direct product of two solvable groups is solvable.*

Proof. It is a particular case of the preceding problem: The direct product $G = K \times B$ defines the natural exact sequence

$$1 \to K \to G \to B \to 1. \qquad \square$$

We return to the monodromy groups of algebraic functions.

13.4 Monodromy groups of algebraic functions

The preceding result implies the main topological fact of Abel's theory:

Theorem 2. *If the monodromy groups of the algebraic functions f and g are solvable, then the monodromy group of the algebraic function $f \circ g$ (whose value at a is $f(g(a))$) is solvable too.*

Proof. Let $\Sigma_g \subset \mathbb{C}$ be the finite set formed by the ramification points of the multi-valued complex algebraic function g and let Σ be the finite set consisting of the ramification points of the multi-valued complex algebraic function $f \circ g$.

We include Σ_g into Σ even if some of the ramification points of g disappear for the function $f \circ g$. The monodromy groups of the algebraic functions g and $f \circ g$ are the representations of the fundamental groups

$$\pi_1(\mathbb{C} \setminus \Sigma_g, *) \quad \text{and} \quad \pi_1(\mathbb{C} \setminus \Sigma, *)$$

by the permutations of the values of the functions g and $f \circ g$, respectively, at the basic point $* \in \mathbb{C} \setminus \Sigma$

Suppose g is n-valued in $\mathbb{C} \setminus \Sigma_g$ and write $z_k(0)$, $1 \leqslant k \leqslant n$, for the values of $g(*)$. When the point a travels from $a(0) = *$ along some continuous path in $\mathbb{C} \setminus \Sigma_g$, these n values of g can be prolonged by continuity, according to the homotopy lifting theory (see pp. 61-73).

Write z_k, $1 \leqslant k \leqslant n$, for the resulting n values over $a(t) \in \mathbb{C} \setminus \Sigma_g$.

We subdivide the values of the function $f \circ g$ at $a(t)$ into n collections: Each collection consists of the r values of $f(z_k(t))$, where the algebraic function f is supposed to be r-valued.

The monodromy group of $f \circ g$, which permutes the rn values of the function $f \circ g$ along some loop in $\mathbb{C} \setminus \Sigma$, has the following structure : The above

n collections of r values are permuted by the monodromy of the function g along the same loop, but the permutations inside each of these collections depend on the homotopy class of that loop in $\mathbb{C} \setminus \Sigma$, rather than in $\mathbb{C} \setminus \Sigma_g$: The same homotopy class in $\mathbb{C} \setminus \Sigma_g$ is represented by different loops in $\mathbb{C} \setminus \Sigma$, since they can make different turns around the new ramification points of the function $f \circ g$, added by f to the old ramification points, which form Σ_g.

Write G and H for the respective monodromy groups of $f \circ g$ and g. The natural inclusion map $i : \mathbb{C} \setminus \Sigma \to \mathbb{C} \setminus \Sigma_g$ induces a homomorphism of the fundamental groups,

$$i_* : \pi_1(\mathbb{C} \setminus \Sigma, *) \to \pi_1(\mathbb{C} \setminus \Sigma_g, *),$$

inducing the homomorphism of the monodromy groups $j : G \to H$. The kernel of this homomorphism j, $\operatorname{Ker} j \hookrightarrow G$, consists of the permutations of the values of $f \circ g$ which preserve each of the n collections of r values, permuting only the values of such a collection between them.

The permutations of the elements inside of each collection are provided by the monodromy of the function f. The permutation inside the kth collection is determined by the monodromy of f along the loop $\{z_k(t)\}$.

Thus, these permutations (associated to different loops in $\mathbb{C} \setminus \Sigma$) belong to the monodromy group of the function f, which we denote by L.

Finally, using the above homomorphism $j : G \to H$, we construct the following exact sequence, which relates the monodromy group G of the complex algebraic function $f \circ g$ to the respective monodromy groups H and L of the functions g and f:

$$1 \to \operatorname{Ker} j \to G \xrightarrow{j} H \to 1, \tag{$*$}$$

where Ker is a subgroup of the direct product L^n of L with itself n times.

If the monodromy group L of f is solvable, then the direct product L^n is solvable and, hence, its subgroup Ker is solvable.

Knowing the solvability of the monodromy group H of g, we conclude from the general results of the problem of page 503 that the exact sequence $(*)$ with solvable K and H implies the solvability of the group G. Theorem 2 is proved. $\square$

Remark. It follows immediately that the monodromy group of any finite combination of radicals and of rational operations, like $\frac{(\sqrt{z}+\sqrt[3]{z+1})^5}{\sqrt[4]{z-1}}$, has a solvable monodromy group. Indeed, the monodromy group of the sum of two algebraic functions, $f + g$, is a subgroup of the direct product of their monodromy groups. It is therefore solvable if both functions have solvable monodromy groups.

The proof is similar for the other operations: Decomposing the finite combination into elementary parts, and applying Theorem 2 many times, we conclude that the monodromy group of the finite combination remains solvable.

Note that in all these reasonings one can replace an algebraic function, whose graph is the union of several algebraic components (say, for the function $\sqrt{z^2} = \pm z$), by any of its components: The monodromy of a component is solvable, if the total monodromy of the function (with additional values) is solvable. This follows from the fact that the subgroups and the quotient groups of solvable groups are solvable.

Thus, in the theory that we are describing, the "representation of a function in terms of radicals" may be understood as the "representation of a function as one of the branches of the combination of radicals".

We have thus studied the topological properties of the representable functions: Their monodromy groups are solvable.

13.5 Topological Proof of Abel Theorem

Main Claim. *The monodromy group of the algebraic function $z(a)$ defined by the equation*

$$z^5 + az + 1 = 0, \tag{6}$$

is not solvable.

Problem. Calculate the monodromy group of the complex algebraic function defined by equation (6).

Solution. The points of ramification are provided by the implicit function theorem. Namely, there is no ramification of the roots (as functions of the parameter a) at the points $a \in \mathbb{C}$ where the derivative in z of the left hand side polynomial of (6) is different from zero. Hence the ramification set, formed by the points $a \in \mathbb{C}$ for which the polynomial (6) has a multiple root, is given by the system of equations

$$z^5 + az + 1 = 0\,, \qquad 5z^4 + a = 0\,.$$

The second equation provides for a the value $-5z^4$. Substituting it into the first equation, we get $4z^5 = 1$, that is, $z = \sqrt[5]{1/4}$. Substituting again into the second equation, we get $a = -5\sqrt[5]{\frac{1}{256}}$.

The ramification set consists therefore of five points a_j, $j = 1, \ldots, 5$, which form a regular pentagon in the plane of the complex numbers a.

To study the monodromy, we choose the basic point $*$ to be $a = 0$. The five values of z at this point are

$$z_k(0) = \sqrt[5]{-1} = e^{\frac{\pi+2k\pi}{5}i}, \quad k = 0, 1, 2, 3, 4.$$

Moving continuously the point $a = 0$ to a ramification point a_j, we produce, at the arrival to the ramification point, a collision of two continuously defined branches, $z_k(t) = z_\ell(t)$.

Therefore, the homotopy class of the loop that connects 0 to a_j, makes a turn around a_j and returns next to $a = 0$ along the arrival path, is represented in the monodromy group by the permutation that transposes the two elements z_k and z_ℓ and leaves invariant the other three values of the function $z(a)$.

In order to arrive to the next point a_{j+1} from 0, we may use the same strategy by turning the above loop at the angle $2\pi/5$, that is, multiplying the values of $a(t)$ along this loop by $\alpha = e^{2\pi i/5}$.

If we divide the values of z by α, we preserve the relation (6) because $\alpha^5 = 1$. So the monodromy along the turned loop $\alpha\, a(t)$ permutes the turned versions z_k/α and z_ℓ/α, leaving invariant the three others.

Turning again and again, we observe that the whole monodromy group of the algebraic function defined by equation (6) is generated by the five transpositions of the neighbours

$$0 \leftrightarrow 1\,, \quad 1 \leftrightarrow 2\,, \quad 2 \leftrightarrow 3\,, \quad 3 \leftrightarrow 4\,, \quad 4 \leftrightarrow 0\,,$$

for the five values z_k. So, the solution of the above problem is given by

Theorem 3. *The monodromy group of the complex algebraic function* $z(a)$ *defined by equation* (6) *is the group of permutations of five elements* $S(5)$ (consisting of 120 permutations).

The unsolvability of this group would imply the impossibility to represent $z(a)$ in the form of a finite combination of radicals and rational operations: The monodromy group of any such combination is solvable, as we have proved in Theorem 2 and the Remark of page 505. Hence, we have to prove the

Theorem 4. *The group of permutations of five elements $S(5)$ is not solvable.*

Proof. If $S(5)$ were solvable, its subgroups would be solvable (see p. 501). Thus, it suffices to prove that its subgroup $S^+(5)$, which consist of the 60 even permutations, is not solvable.

The geometry of the dodecahedron is very helpful for the algebraic investigation of the groups $S(5)$ and $S^+(5)$. On pp. 51-52 we have used the five cubes of Kepler inscribed into the dodecahedron to study the group $S^+(5)$ of the 60 even permutations of 5 elements, where we have shown that it is naturally isomorphic to the group of the 60 rotational symmetries of the dodecahedron. To prove Theorem 4 we need the following fact.

Proposition. *The group G of the* 60 *rotations of the dodecahedron has no invariant subgroup (except the trivial subgroups e and G).*

We present here a slightly different proof to that given on pp. 51-52.

Proof of the Proposition. Any rotation of a dodecahedron preserves either a face, or an edge, or a vertex (or it is the identity rotation e). Suppose G has an invariant subgroup H. The invariance of H implies that together with any non-trivial rotation preserving a face the subgroup H contains all the $4 \cdot 6 = 24$ non-trivial rotations that preserve a face, since there exist rotations sending one face to any other face. Similarly, together with any non-trivial rotation preserving a vertex the subgroup H must contain all the 20 such rotations, and together with any non-trivial edge preserving rotation – all the 15 such rotations (there are rotations sending one vertex to any other, and rotations sending one edge to any other).

In consequence, the number of elements of the invariant subgroup H is the sum of 1 (for e) plus the sum of some of the three numbers $\{24, 20, 15\}$. There are 8 possible values of such a sum:

$$1,\ 16,\ 21,\ 25,\ 36,\ 40,\ 45,\ 60.$$

Only 1 and 60 are divisors of the total number $|G| = 60$ of rotations. Any invariant subgroup contains therefore either only one element e or all the 60 rotations forming G. The proposition is proved. $\square$

Now, the group $S^+(5)$ of the even permutations of 5 objects has no non-trivial invariant subgroup, since the group of rotations of the dodecahedron permutes the five Kepler cubes inscribed into the dodecahedron and is, hence, isomorphic to the unique subgroup $S^+(5)$ of 60 elements in the group of permutations $S(5)$. Theorem 4 is proved. $\square$

Corollary. *Equation* (6) *cannot be solved in terms of radicals and of finite combinations of rational operations. Abel's Theorem is proved.*

13.6 Topological Imposibility Theorems

The impossibility to solve (in the same sense) the general equation of degree 5 follows from the above corollary. In fact, our arguments provide more:

Claim. *Any equation whose monodromy group contains* $S^+(5)$ *cannot be solved in terms of radicals and of finite combinations of rational operations.*

For example, the generic algebraic equations of degree $n \geqslant 5$ are unsolvable in radicals. The monodromy groups of such equations are $S(n) \supset S(5)$.

Our proof shows that the unsolvability reason is *topological* and, in fact, we have proved the following generalisation of the original Abel theorem.

Topological Unsolvability Theorem. *No complex algebraic function topologically equivalent to the solution* $z(a)$ *of equation* (6)*, or to the solution* $z(a, b, c, d, e)$ *of equation*

$$z^5 + az^4 + bz^3 + cz^2 + dz + e = 0$$

(or to the solutions of similar equations of higher degrees) can be represented in the form of a finite combination of radicals and of rational operations and of any univalent function.

As we have seen, this analytic-topological result, which seems to belong to logic and to algebra, is a manifestation of the profound unity of all branches of mathematics. We have used the Kepler cubes from stereometry, the monodromy of ramified coverings from topology, the exact sequences from algebra, and even number theory is closely related to Abel's theorem by the Galois generalisation that proves the unsolvability for some individual values of the coefficients.

There are reasons to believe that many impossibility results of mathematics have topological nature and should be proved in the form of a topological impossibility statement, as we have done above proving Abel's theorem.

Divergence. Topological impossibility occurs in many problems of geometry and calculus. For example, the divergence of the Taylor series of the function $\arctan x$ for $|x| > 1$ has a topological reason: a pole of the derivative at the complex point $x = i$, invisible in the real x-axis (rather than the big zise of the Taylor coefficients).

Three Body Problem. Poincaré theorem on the absence of any additional analytic first integral, independent of the classical ones, should be generalised by stating the absence of any first integral in any problem whose phase portrait is topologically similar to that of Poincaré's problem. That theorem, alas, has never been published.

Poincaré's theorem is essentially topological. Namely, investigating the generation of periodic orbits by the resonances of the approximated problem of independent Keplerian planets, Poincaré discovered the birth of numerous non-degenerate isolated (perturbed) periodic orbits. However, he observed that in presence of an additional first integral such orbits should not be isolated, but ought to form continuous families, absent in the case of his study. Consequently, there is no new first integral in this case.

Strangely, this clear topological reasoning is hidden in the Poincaré description of his theorem: His proof is correct, but it provides less information on the nature of the phenomena than that he has informally discovered when prepared his proof.

The same remark is also applicable to the theorems on the impossibility of solution "in quadratures" of differential equations (using the integration operation).

Most impossibility results in mathematics are due to the topological complexity of the behaviour of the systems that one wishes to describe by using simple models for which such complicate behaviour is impossible. The description of those complexities is sometimes more important than the impossibility theorem itself.

Elliptic Integrals. Write $t(X)$ for the value of the elliptic integral

$$t(X) = \int_{X_0}^{X} \frac{dx}{\sqrt{x^3 + ax + b}},$$

considered as a (multivalued) complex function of complex argument X, and write $X = F(t)$ for the inverse doubly-periodic meromorphic (elliptic) function (see p. 154).

Suppose that the complex numbers a and b are generic (such that the cubic polynomial of the denominator has no multiple roots).

In his lectures of 1964, Arnold attributed to Abel the following result.

Theorem. *Neither the elliptic integral $t(X)$ nor the elliptic function $X(t)$ is topologically equivalent to any elementary function.*

Topological equivalence. Two functions $f : A \to B$, $\widetilde{f} : \widetilde{A} \to \widetilde{B}$ are *topologically equivalent*, if there exist two homeomorphisms $h : A \to \widetilde{A}$ and $k : B \to \widetilde{B}$, transforming f into $\widetilde{f}$. That is, $\widetilde{f}(a) = k(f(h^{-1}(a)))$ for any a in $\widetilde{A}$:

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle k} \\ \widetilde{A} & \xrightarrow{\widetilde{f}} & \widetilde{B}\,. \end{array}$$

In this theorem, for the meromorphic doubly-periodic function $X = F(t)$, we have $A = \mathbb{C} = \{t\}$, $B = \mathbb{C}\mathrm{P}^1 \ni X$ and $f = F$.

This theorem shows that the impossibility to integrate elliptic or other abelian integrals by elementary functions, along non-zero genus algebraic curves, has topological reasons. One can express this by saying that

The topological properties of the elementary functions are too restrictive to include the topological complications of elliptic or abelian integrals.

Unfortunately, neither the proof of this "Abel theorem" nor that of its generalisations to other Abelian integrals and functions have been published yet, in spite of the fact that it was suggested to the listeners of the 1963-1964 lectures, being then Moscow high-school students.

We hope that Arnold's topological proof of Abel's Theorem will open the way to many topological impossibility results. Some topological ideas of Arnold's lectures were developed by A.G. Khovanskii, who has proved some new results on the unsolvability of differential equations. Unfortunately, the topological insolvability proofs are still missing in his theory (as well as in the Poincaré theory on the absence of the holomorphic first integral and in many other unsolvability problems of differential equations theory).

Chapter 14

Newton Polyhedra: Geometry of formulae

14.1 Neighbouring volumes of submanifolds

We start with some very natural and simple problems.

Problem. Find the area of the domain formed by the points in Euclidean plane, whose distances to a given convex polygon of that plane are smaller than ε.

Solution. Fig. 14.1 explains that it is a second degree polynomial,

$$S(\varepsilon) = S_0 + L_0\varepsilon + \pi\varepsilon^2,$$

where S_0 is the area of the polygon and L_0 is its perimeter.

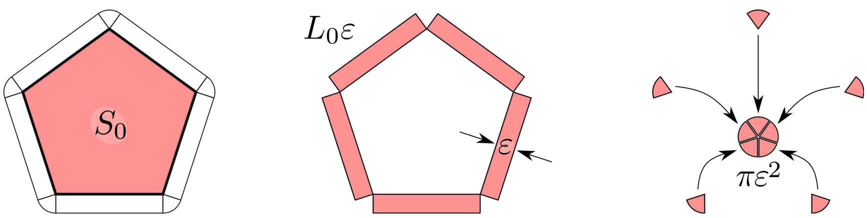


Figure 14.1: Proof of the polynomialness of the neighbouring area function $S(\varepsilon)$.

Problem. Find the length excess of the curve parallel to the Earth equator, but situated one meter higher.

SOLUTION. The length is equal to $L(\varepsilon) = dS/d\varepsilon$, hence $dL/d\varepsilon = d^2S/d\varepsilon^2 = 2\pi$, and thus the excess is 628 cm.

PROBLEM. Find the volume of the domain formed by the points in Euclidean 3-space, whose distances to a given convex polyhedron are smaller than ε.

SOLUTION. The volume is a polynomial of degree 3 in ε,

$$V(\varepsilon) = V_0 + S_0\varepsilon + L_0\varepsilon^2 + (4/3)\pi\varepsilon^3,$$

where the coefficients are
$V_0 =$ the volume of the polyhedron;
$S_0 =$ the area of the polyhedral surface;
$L_0 =$ the "length" $c \cdot \ell_0$;
where ℓ_0 is the average of the length of the orthogonal projection of the polyhedron to a straight line, and c is a universal coefficient independent on the polyhedron, like $(4/3)\pi$ in the cubic term, that we shall compute in a while.

PROBLEM. Find the volume of the domain formed by the points in Euclidean n-space, whose distances to a given convex smooth body are smaller than ε.

ANSWER. The volume is a degree n polynomial in ε,

$$V(\varepsilon) = V_0 + V_1\varepsilon + V_2\varepsilon^2 + \ldots + V_n\varepsilon^n. \tag{1}$$

Up to some universal factor c_j that does not depend on the body, the coefficient V_j is the integral, along the body boundary, of the basic symmetric function of the principal curvatures of the hypersurface bounding the body:

$$V_j = c_j \int\limits_{x \in M^m} K_{j-1}(x)\,dx\,, \qquad \text{where}$$

$$K_{j-1}(x) = \sigma_{j-1}\left(k_1(x), k_2(x), \ldots, k_m(x)\right), \qquad m = n-1. \tag{2}$$

The principal curvatures of this smooth boundary hypersurface are supposed here to be positive (the second quadratic form being defined by the distance from the tangent plane, positive inside the body).

The basic symmetric functions $\sigma_j(a)$ of the arguments $a_1, a_2, \ldots, a_m$ are defined by the Vieta formula

$$\prod_{j=1}^{m}(t + a_j) = t^m + \sigma_1(a)t^{m-1} + \sigma_2(a)t^{m-2} + \ldots + \sigma_m(a);$$

that is, $\sigma_1 = a_1 + a_2 + \ldots + a_m$, $\sigma_2 = a_1a_2 + a_1a_3 + \ldots + a_{m-1}a_m, \ldots$, $\sigma_m = a_1a_2\ldots a_m$ (and $\sigma_0(a) \equiv 1$ by definition).

Remark. The integral V_j of the symmetric function K_{j-1} along the boundary hypersurface is proportional (with the proportionality coefficient independent of the body) to the $(n-j)$-dimensional volume ℓ_{n-j} of the orthogonal projection of the body to the $(n-j)$-dimensional subspaces of Euclidean space $\mathbb{R}^n$, averaged along all the projections.

A usual method to determine universal coefficients that describe characteristic properties of geometric objects is to use simple examples in which those characteristic properties are known (or easy to determine). Then the universal coefficients involved in that example can be computed. We shall determine the first three coefficients c_j using the ball in Euclidean 3-space.

Example. For the ball of radius R in Euclidean 3-space the polynomial (1) takes the form

$$V(\varepsilon) = V_0 + V_1\varepsilon + V_2\varepsilon^2 + V_3\varepsilon^3 = (4\pi/3)(R+\varepsilon)^3,$$

hence $V_0 = (4\pi/3)R^3$, $V_1 = 4\pi R^2$, $V_2 = 4\pi R$, $V_3 = 4\pi/3$.

Since the principal curvatures are $k_1 = k_2 = 1/R$, we get $K_0 = 1$, $K_1 = 2/R$, $K_2 = 1/R^2$. We denote by ω a variable point on the sphere of radius R and, in terms of ω we get the coefficients V_j :

$$V_1 = \oiint c_1K_0 d\omega = 4\pi R^2 c_1;$$

$$V_2 = \oiint c_2K_1 d\omega = 4\pi R^2(2/R)c_2 = 8\pi R c_2;$$

$$V_3 = \oiint c_3K_2 d\omega = 4\pi R^2(1/R^2)c_3 = 4\pi c_3.$$

Now we will use these values of V_j that we have obtained for the sphere to find the universal constants: $c_1 = 1$, $c_2 = 1/2$, $c_3 = 1/3$ for the expression of the coefficients V_j as the integrals of c_jK_{j-1}.

The area and length of the projections of the sphere of radius R are $\ell_2 = \pi R^2$ and $\ell_1 = 2R$. Hence, for the sphere we get

$$V_1 = 4\ell_2, \qquad V_2 = 2\pi\ell_1,$$

where the coefficients 4 and 2π are independent of the body.

Example. For the cube of edge of length R, we have

$$V(\varepsilon) = R^3 + 6R^2\varepsilon + 12R(\pi/4)\varepsilon^2 + 8(4\pi/3)\varepsilon^3/8,$$

that is, $V_1 = 6R^2$, $V_2 = 3\pi R$, $V_3 = 4\pi/3$.

Thus, the universal coefficients calculated above provide the averaged area of the projection of the cube to a random plane: $l_2 = V_1/4 = \frac{3}{2}R^2$. Here the minimal projected area is R^2 and the maximal one (hexagonal) is $\sqrt{3}R^2$.

Similarly, the above universal coefficient provide us the averaged length of the orthogonal projections of the cube of edge of length R to a line: $\ell_1 = V_2/(2\pi) = (3/2)R$ – in this case, the minimal projected length is R and the maximal one is $\sqrt{3}R$.

14.2 Proof of neighbouring volume formula with principal curvatures

Denote the boundary hypersurface of our convex body by M^m, $m = n-1$, and consider the map $F : M^m \times [0, \varepsilon] \to \mathbb{R}^n$, sending the point $x \in M$ and the distance $t \in [0, \varepsilon)$ to the point $F(x,t) = x + t\nu(x)$ of $\mathbb{R}^n$, where $\nu(x)$ is the exterior unit normal vector to M at x.

The image of the map F covers just the exterior domain where the distance to M^m is smaller than ε, and thus

$$V(\varepsilon) = V(0) + \int\limits_{M^m \times [0,\varepsilon)} \det(F_{*(x,t)}) dx \wedge dt,$$

where dx is the m-dimensional volume element on M^m.

To calculate the Jacobi determinant $\det(F_{*(x,t)})$, note that $\partial F/\partial t = \nu(x)$ is the unit vector orthogonal to $\partial F/\partial x$, whence this determinant is equal to the determinant of the $m \times m$ matrix $1 + t(\partial\nu/\partial x)$ of the operator acting on the tangent m-space at x, $T_x M^m$.

Taking into account the definition of the principal curvatures and our orientation convention, which makes them positive for the boundary hypersurface M^m, we get the

Lemma. *The principal curvatures of the hypersurface M^m at its point x are the eigenvalues of the linear operator*

$$\partial\nu/\partial x : T_x M^m \to T_x M^m.$$

So, the determinant of the $m \times m$ matrix $1 + t(\partial\nu/\partial x)$ is given in terms of the principal curvatures by the formula

$$\det(1 + t(\partial\nu/\partial x)) = \prod_{j=1}^{m}(1 + tk_j) = 1 + t\sigma_1 + t^2\sigma_2 + \ldots + t^m\sigma_m,$$

where $\sigma_j = \sigma_j\,(k_1(x), \ldots, k_m(x))$ is the jth basic symmetric function.

Integrating from 0 to ε along the t variable we get

$$\int_0^\varepsilon \det(1 + t(\partial\nu/\partial x))dt = \sum_{j=1}^{n} \frac{\varepsilon^j}{j}\sigma_{j-1}.$$

Thus, the coefficients of the neighbouring volume polynomial (1) are

$$V_j = \frac{1}{j} \int\limits_{x \in M^m} K_{j-1}(x)dx$$

and, hence, we have calculated the promised universal constants: $c_j = 1/j$.

14.3 Proof of the formula for the averaged volumes of projections

Consider the manifold E of the j-dimensional affine subspaces of the Euclidean n-space $\mathbb{R}^n$. We equip it with the natural Riemannian structure that is invariant under the isometries of the affine Euclidean space and that provides the corresponding invariant measure on E.

Associate to the smooth submanifold M^m of $\mathbb{R}^n$ (which is the boundary of a convex body, $m = n - 1$) the submanifold X of E that consist of the j-dimensional affine subspaces that are tangent somewhere to M^m. Any such j-dimensional space, tangent to M at some point, is contained in the tangent affine m-dimensional space of M^m at that point.

The manifold X is a compact smooth submanifold in E, fibred smoothly over the convex hypersurface M^m.

We can compute the volume of X in two ways: Either integrating first along M^m for a fixed affine subspace direction and then integrating the result along the choices of the direction of projection, or integrating first along the choices of the affine tangent j-subspaces of M^m for a fixed point and then integrating along M^m.

Both orders of integration provide equal double integrals by the Fubini theorem, but they have different geometric contents.

Fixing the direction of the projection to a j-space, the value of the integral is the j-volume of the projection of the boundary hypersurface M^m to this j-space. The final integration along the different projections provide the averaged j-volume, ℓ_j.

The result of the integration at a fixed element of the hypersurface M^m depends on the principal curvatures of M^m at this place in such a way that the higher values of the curvatures provide a larger contribution to the volume of X in E.

The usual calculations of the Jacobian provide just the symmetric function K_{j-1} of the curvatures as the coefficient defining the ratio of the volume of the image of the element in X (measured by the metrics of E) to the original volume of the element in M producing this infinitesimal part of X.

Consequently, the equality of the results of the integration in both orders implies that the integral of the symmetric function K_{j-1} of the principal curvatures along the convex m-surface M^m is equal to the averaged volume of the j-dimensional projections.

We shall not compute here the universal constants of proportionality provided by these integrals. We prefer to calculate these constants by considering simple examples, like the spheres and the cubes above.

The resulting proportionality coefficients are the ratios of the volumes of the unit spheres, and hence they have explicit expressions in terms of the Euler gamma function.

Remark. All the preceding theory has nice extensions to the case of non-convex hypersurfaces and also to the case of subvarieties of smaller dimension, including even the space curves (counting the volumes of the projections with appropriate multiplicities that depend on the number of components of the projected section).

For any surface of genus g in $\mathbb{R}^3$, the term of highest degree $(4\pi/3)\varepsilon^3$ of the cubic polynomial of the 3rd problem above, becomes proportional to its Euler characteristic, being $(2/3)\pi(2-2g)\varepsilon^3$ due to the equality to $1-g$ of the degree of the Gauss map.

Thus, for the torus in $\mathbb{R}^3$ the cubic term in formula (1) disappears:

$$V(\varepsilon) = V_0 + S_0\varepsilon + L_0\varepsilon^2.$$

This independence of the coefficient of the cubic term from the details of the surface reflects thus the topologic invariants of the surface. This fact has lead researchers on this subject (called "tube volumes theory") to study other similar integrals involved in these geometric problems: Are these integrals really dependent on the details of the submanifold or they represent its topologic properties (like the Gaussian curvature integral)?

This question leads to the study of the variations of such geometric integrals: The problem was to find those integrals whose variations vanish automatically whatever would be the small variation of the embedded submanifold in Euclidean space.

Such integrals where finally discovered and are called today the cohomologic characteristic classes of the original manifold embedded in Euclidean space. The Euler class (represented by the integral of the Gaussian curvature) is their first representative.

All the complicated theory of the characteristic classes is a byproduct of the practical problem of computing the volumes of tubes, invented by the statistics department people, who, being unable to compute the volumes, asked Herman Weyl to help them. Weyl provided then the basic material for his further brilliant works on this subject and for the discovery of the characteristic classes.

Practical applications are often stimulating the best theoretical works.

14.4 Discrete mixed volumes

The theory of neighbouring volumes has an astonishing number-theoretic counterpart.

Consider a convex polyhedron with integer vertices in Euclidean space $\mathbb{R}^n$ with a fixed lattice of integer vectors $\mathbb{Z}^n \subset \mathbb{R}^n$. Let the polyhedron be defined by a finite system of linear inequalities, $\langle a_k, x\rangle \leq 1$.

Instead of the volumes of the preceding theory, in the number theoretical version the object of study will be the number of integral points.

Notation. Given a positive integer t, write $N(t)$ for the number of integer lattice points x verifying the inequalities $\langle a_k, x\rangle \leq t$.

Theorem 1. *The function N is a polynomial of degree n*

$$N(t) = N_0 + N_1 t + \ldots + N_n t^n.$$

Thus, this polynomial counts the integral points inside the polyhedron $\{x : \langle a_k, x\rangle \leq t\}$, homothetic to the original one.

Theorem 2. *The number $N^+(t)$ of the lattice points strictly inside the homothetic polyhedron* (verifying the strict inequalities $\langle a_k, x\rangle < t$) *is a polynomial of degree n, namely*

$$N^+(t) = (-1)^n N(-t).$$

The integration of the coefficients of these polynomials in terms of the arithmetic properties of the faces of different dimensions of the original polyhedron and of its "integral angles", provides interesting arithmetic extensions of the curvatures. But the resulting number-theoretic invariants are sometimes difficult to handle.

Example. (Fig. 14.2) For the plane triangle with vertices $(0, 0)$, $(1, 0)$ and $(0, 1)$ we get

$$N(1) = 3, \quad N(2) = 6, \quad N(3) = 10, \quad N(4) = 15, \quad \ldots \quad \text{and}$$

$$N^+(1) = 0, \quad N^+(2) = 0, \quad N^+(3) = 1, \quad N^+(4) = 3, \quad \ldots$$

which are the respective values of the polynomials

$$N(t) = (t + 1)(t + 2)/2 \quad \text{and} \quad N^+(t) = (t - 1)(t - 2)/2,$$

confirming both theorems.

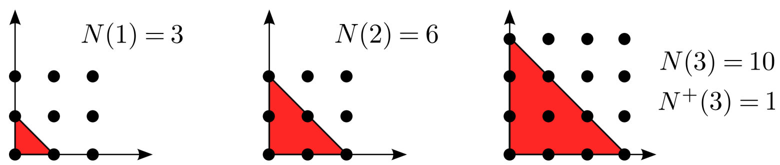


Figure 14.2: Integral lattice points inside homothetic triangles.

The proof of Theorem 1 is essentially contained in this example. One can subdivide the initial polyhedron into simplicial pyramids* and then, counting integral points in the homothetic pyramids, one adds the results. To take into account the points of the subdividing hyperplanes, which are counted

*This fact is related to the celebrated Hilbert theorem on finite bases of ring theory.

twice, one apply the $(n-1)$-dimensional version of Theorem 1, and prove the whole theorem by induction on the parameter n.

The counting of the points in the pyramid may be also done inductively using the finite differences $N(t+1)-N(t)$, which count the number of integral points on the $(n-1)$-dimensional hyperplanes, provided by the study of the $(n-1)$-dimensional case. Knowing the difference to be a polynomial of degree $n-1$, we conclude that N is a polynomial of degree n.

We leave to the reader the formal details of this proof which provides also the second theorem, interpreting the difference $N(t) - N^+(t)$ as the number of integral points on the $(n-1)$-dimensional boundary.

The interpretation of the coefficients of the polynomial (1) as volumes of projections should be also extended to the arithmetic situation, but we have not seen any written version of it.

The distribution of integral points in convex polyhedra is an important part of "Newton polyhedra theory" (generalising the "Newton parallelogram" providing the Puiseux series), of Euler's theory of graded algebras (exposed in his "Introduction to calculus"), of Hilbert's theorem on "finite bases of ideals" and of the "Gröbner basis theory", extending the Hilbert "theological theorem" to applied mathematics problems of asymptotic expansions of the solutions of algebraic and differential equations.

This modern theory of "geometry of formulas" can be studied in the elementary textbook [89] by A.G. Khovansky and in his articles [87, 88].

In all these applications the volumes of the convex bodies are mostly replaced by the more general notion of the Minkovsky "mixed volumes" on which we shall say few words.

The "Minkovsky sum" $X + Y$ of two convex bodies in a vector-space consists of all the sums $x+y$ of the points of the bodies. One defines similarly the multiplication of a convex body by a scalar cX ($c \in \mathbb{R}$). Identifying the bodies obtained one from other by a translation, one obtains a vector space structure on the set of formal linear combinations of convex bodies of an affine space. Write B for this vector space.

The volume of a convex body in $\mathbb{R}^n$ is considered in this theory as "a homogeneous polynomial" of degree n.

Definition. The *mixed volume* $F(X_1, X_2, \ldots, X_n)$ of n bodies in the n-space $\mathbb{R}^n$ is a symmetric multilinear function $F : (B \times \ldots \times B = B^n) \to \mathbb{R}$, which coincides with the volume f on the diagonal.

Example. $F(X,Y) = \frac{1}{2}[f(X+Y) - f(X) - f(Y)]$.

Problem. Compute the mixed volume of two triangles in the plane, having vertices $((0,0),\ (0,a),\ (b,0))$ and $((0,0),\ (0,c),\ (d,0))$.

Answer. $\max(ad, bc)/2$.

The mixed volumes are very helpful to represent complicated answers in such problems of mathematics as the computation of the multiplicity of a multiple root of a system of algebraic equations and of the coefficients of the irreducible representations of a group in the decomposition of the tensor product of two of them (the "Clebsh-Gordan coefficients").

They also provide the Betti numbers (and the Hodge numbers*) of the projective algebraic manifolds defined by systems of algebraic equations, in terms of the stereometry of the points in the Newton space that represent the monomials having nonzero coefficients in the equation.

14.5 The Newton polyhedra

Consider an algebraic curve M defined by a polynomial equation $f(x, y) = 0$, where $f = \sum f_{u,v} x^u y^v$ with generic complex coefficients.

Definition. The convex hull of the points $(u, v) \in \mathbb{Z}^2 \subset \mathbb{R}^2$ that correspond to the non vanishing coefficients $f_{u,v}$, is a convex polygon P called *the Newton polygon* of that polynomial equation.

It turns out that the genus $g(M)$ of this curve is equal to the number of the strictly interior integer points of the convex polygon P.

So for the elliptic curve $f = 0$, where

$$f = y^2 + x^3 + ax^2 + bx + c,$$

the Newton polygon P has one interior integral point, which provides thus the genus $g = 1$ – Fig. 14.3.

For the "hyperelliptic curve" defined by the equation

$$y^2 = x^k + a_1 x^{k-1} + \ldots + a_k,$$

the number of integral points strictly inside the triangle P is g for $k = 2g+1$ or $2g + 2$ (Fig. 14.4). Hence, the genus of the hyperelliptic curve is

$$g = \left[\frac{k-1}{2}\right],$$

*Although the Hodge numbers are not defined nor used in this book, below, we discuss briefly some idea about their meaning.

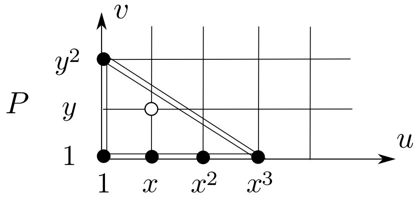


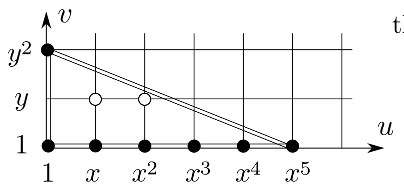
Figure 14.3: The newton polygon P of an elliptic curve.

k	$1,2$	$3,4$	$5,6$	$7,8$	$\dots$
g	0	1	2	3	$\dots$

.

$$xy \mapsto \int \frac{dx}{y}$$

$$x^2y \mapsto \int \frac{xdx}{y}$$



the two interior points xy and x^2y provide the genus $g = 2$

Figure 14.4: The Newton polygon of a hyperelliptic curve of degree 5, and the basic Abelian integrals along this curve.

The interior points of the Newton polygon of an algebraic curve of genus g do correspond to the g basic Abelian integrals along that algebraic curve. In the case of the hyperelliptic curve those Abelian integrals are

$$\int \frac{x^j dx}{y}\ , \quad j = 0, 1, \dots, g-1.$$

For a more general algebraic curve the description of the forms in terms of the interior points of its Newton polygon is similar. Namely, the interior points (α, β) of the Newton polygon of the algebraic curve $f = 0$ determine the forms

$$x^{\alpha-1} y^{\beta-1} \frac{dx \wedge dy}{df},$$

which have not singularities and form a base of the space of holomorphic forms on that curve.

We have to precise what means the “quotient” $\frac{dx\wedge dy}{df}$. This “quotient” is a formal notation for the so-called *Gelfand-Leray* 1*-form*,

$$\alpha := \frac{dx \wedge dy}{df},$$

which is the 1-form on the curve $f = 0$, defined by the condition

$$df \wedge \alpha = dx \wedge dy.$$

On the plane, such α is not unique because it is defined up to a term $g \cdot df$. But the restriction of α to the curve $f = 0$ is well defined, since df vanishes along the curves $f = \text{const}$.

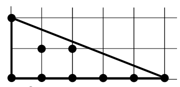
For a smooth curve, at least one of the coordinates x or y can be taken as a local coordinate along the curve. So, using that $df = f_x dx + f_y dy$ and supposing $\alpha = *dx$ or $\alpha = *dy$, the condition $df \wedge \alpha = dx \wedge dy$ leads to the respective equalities

$$\frac{dx \wedge dy}{df} = -\frac{1}{f_y}\,dx \quad \text{or} \quad \frac{dx \wedge dy}{df} = \frac{1}{f_x}\,dy,$$

which are two different expressions of the Gelfand-Leray 1-form in the respective coordinate systems on the curve $f = 0$.

Consider the hyperelliptic curve $f = 0$, where

$$f(x, y) = -y^2 + x^5 + a_4 x^4 + \ldots + a_1 x + a_0.$$

Its Newton polygon, , has the two interior points $(1, 1)$ and $(2, 1)$. The extension of this curve to the projective plane has a singularity at the point where the curve meets the line at infinity.

In the set $(\mathbb{C} \setminus \{0\})^2$, the curve $f = 0$ is invariant under multiplication of f by a monomial. For example, the curve $x^2 y^3 f = 0$ coincides with the curve $f = 0$ on $(\mathbb{C} \setminus \{0\})^2$. Moreover, its Newton polygon is a translation of that of f along the vector $(2, 3)$, having then the same number of interior points.

The coordinate change $X = x^a y^b$, $Y = x^c y^d$, where $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}(2, \mathbb{Z})$, is invertible and the inverse change of coordinates is given by a similar formula but the exponents are provided by the inverse matrix.

The stereometry of the Newton polyhedra provides also similar explicit stereometric formulae for the "mixed Hodge numbers" of singular or non closed algebraic varieties.

The mixed Hodge structure associates to each homology class of an algebraic variety its ancestors in the homology classes of some smooth algebraic varieties.

Say, for the variety consisting of three elliptic curves that intersect each other, the real one-dimensional triangular cycle connecting the three intersection points is not homologous to any combination of the parallels and meridians of the elliptic curves. In this example the triangle is a well defined cycle, modulo the combinations of parallels and meridians. The mixed Hodge numbers count such cycles of different origins in singular varieties (or in non closed varieties).

These pseudo-topological integers and structures are strangely related rather to the algebraic structure of the variety than to its topology. They are respected by the algebraic and semi-algebraic maps (say, by the linear projections of algebraic subvarieties of vector-spaces), but there exist holomorphic (non algebraic) diffeomorphisms that change the values of the mixed Hodge numbers.

We observe here an important difference between algebra and geometry. The integer-valued characteristics (like the genus of an algebraic curve) have in most cases the property of topological invariance or at least the invariance under diffeomorphisms. But the mixed Hodge structures and numbers have no topological explanations, providing a lot of difficulties to algebraic geometry, which would otherwise be a part of the geometry science.

It is astonishing that these mixed Hodge numbers had been already computed by O.A. Oleinik in her studies of the real algebraic geometry for the 16th Hilbert problem, which provide the upper bounds for the Betti numbers of the complex projective algebraic varieties in terms of the degrees of their equations*.

Oleinik's theory described the bounding numbers by long chains of recurrent relations. But V. Arnold recognised in these chains the complicated expressions of the volumes of convex polyhedra and of the numbers of their interior integral points, in terms of the coordinates of the vertices.

Replacing the complicated recurrent formulas by the simple geometric words "volume of the polyhedron with such vertices", or "mixed volume", or "number of interior integer points", one gets a nicer description of the results and one can work with these volumes and numbers of integral points faster than with their (existent) explicit expressions in terms of the coordinates of the vertices.

As far as we know, the formulae presented above for the calculation of

*These bounds are strangely attributed in the West to Thom and Milnor who published their weaker results 15 years later than Oleinik.

volumes in terms of the curvatures have not been yet extended to the mixed volume theory of smooth convex bodies.

The idea of Newton was that a polynomial or even a series

$$f(x_1, x_2, \ldots, x_n) = \sum_{u \in \mathbb{Z}^n} \widehat{f}_u x^u$$

(where the monomial x^u is $x_1^{u_1} \ldots x_n^{u_n}$) defines a new function $\widehat{f} : \mathbb{Z}^n \to \mathbb{C}$ of the integral point u. This new function is called today "the Fourier transform of f", with slightly different notations.

The *support* of f consists of the points u at which the coefficient $\widehat{f}(u)$ of the polynomial f is not zero. Thus, the Newton polyhedron P of f is the convex hull of the *support* of f.

The observations of Newton on these polyhedra were even considered by him as his *main contribution to mathematics.* He attributed the derivatives and the integrals, as well as their interrelations, and even the gravitation theory of the Kepler laws, to his predecessors including Archimedes, Barrow, the early Egyptian mathematician Thot (proclaimed to be a god by the Pharaoh) and others. More details on the Newton predecessors, including Thot, can be read in [24, 33, 34].

This main Newton observation relates the geometry of the Newton polyhedron P to the asymptotic behaviour of the variety $f = 0$ at the origin, as it is shown in Fig. 14.5 for the algebraic curve $Ay^5 + Bxy^2 + Cx^6 = 0$.

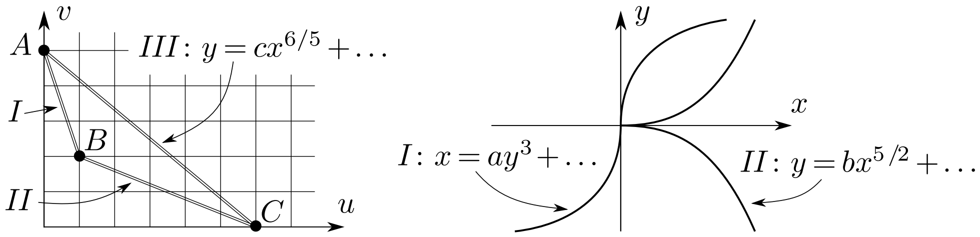


Figure 14.5: The Newton asymptotics of the two branches (I, II) of the algebraic curve $Ay^5 + Bxy^2 + Cx^6 = 0$ at $(0, 0)$: The side $Ay^5 + Bxy^2 = 0$ provides the branch $I : x = -A/B\ y^3 + \ldots$.

Namely, for the case $n = 2$ of a plane curve $f(x, y) = 0$, Newton suggested to replace the polynomial f by a smaller sum $\widetilde{f}$ of its terms: One takes only those terms of f that are supported by one of the sides of the Newton polygon of f that is visible from the origin of the (u, v)-plane.

The statement is that near the point $x = y = 0$, the simpler curve $\widetilde{f} = 0$ provides a good approximation of one of the branches of the curve $f = 0$, say $y = Cx^{p/q}$. Next, substituting the expression $y = Cx^{p/q} + z$ in the equation $f(x, y) = 0$, Newton obtained a new algebraic equation in x and z that provides a new polygon and an approximated formula for z. The recursive repetition such substitutions led Newton to the series expansion $y = \sum C_k x^{p_k/q_k}$ of the corresponding branch of the curve $f(x, y) = 0$ near the origin. This Newton series is called at present "sèrie de Puiseux".

The convergence of these series for sufficiently small $| x |$ follows from the Taylor series convergence. Indeed, the function y becomes a holomorphic function of the variable $x^{1/r}$ for a suitable integer r. This follows from the fact that the multivalued *algebraic* function $y(x)$ has only a finite number of values. It returns therefore to its initial value after a suitable number r of turns of the complex variable x around the ramification point $x = 0$.

This reasoning shows that all the fractions p_k/q_k of the above series have a common denominator r.

Similarly, the sides of the Newton polygon which are visible from the infinity point ($u = v = \infty$) provide the asymptotic expansions of the infinite branches of the algebraic curve $f(x, y) = 0$.

So the interior points of the Newton polygon contribute nothing to both asymptotics. The disregard of those terms of the polynomials and of the series that correspond to the interior points of the Newton polygons and Newton polyhedrons, is a strange method of many physicists to delete some terms from their formulas, even in the cases when other terms are preserved in the local study, in spite of the fact that they are smaller than the neglected ones. In such cases the physicists refer to the preserved smaller terms, claiming that they are "physically important", the neglected ones "having less physical meaning". In fact all this "physical unimportance" is the smallness in the Newton filtrations of the ring of functions related to the Newton polygon (or polyhedron).

The extensions to $n > 2$ of this theory are more long to be explained, but the same principle is used. In the case $n = 2$, different branches of the curve corresponded to the different sides of the Newton polygon P. In the cases $n > 2$ the polygon becomes a polyhedral surface which has a complicated net of faces of different dimensions. These stereometric characteristics of the support of a polynomial f of n variables contribute to the particular asymptotic expansions at the corresponding horn-like domains of the neighbourhood of the singular point 0 of the variety $f = 0$ in $\mathbb{C}^n$.

The *fewnomials* are those polynomials f which have, in some coordinate system, few monomials. A particular corollary of this bound of the complexity of the defining formula is the topological simplicity of the real variety $f = 0$. Thus, a real polynomial in one variable

$$f(x) = x^n + a_1 x^{n-1} + \ldots + a_n$$

cannot have many real roots if it has few nonzero coefficients.

Indeed, the Descartes rule claims that the number of positive roots cannot exceed the number of sign changes in the sequence of the coefficients.

So, the number of positive roots cannot exceed the number of the nonzero coefficients, which bounds also the number of the negative roots.

The theory of fewnomials extends this corollary of the Descartes rule to real polynomials in many variables: The smallness of the number of the vertices of the Newton polyhedron P implies the upper bounds for the topological invariants of the corresponding real algebraic varieties. For instance, for the number of their connected components, for the genus of a surface, for the Betti numbers of all dimensions and for the Euler characteristic of the varieties defined by the fewnomials. Similar bounds are conjecturally true even for the solutions of the differential equations defined by short formulas.

The 16-th Hilbert problem suggests such bounds for the topological objects defined by the differential equations whose coefficients are polynomials of small degrees, but even his question on the existence of a common upper bound for the numbers of the limit cycles of arbitrary quadratic vector field in the plane $\mathbb{R}^2$ remains open. His differential equations are

$$\frac{dx}{dt} = P(x, y), \qquad \frac{dy}{dt} = Q(x, y),$$

where P and Q are polynomials of degree $\deg P = \deg Q = 2$.

It is unknown whether the number of limit cycles of all such equations is bounded by a common constant, independent on the coefficients of the polynomials. It is unknown even for the case in which the polynomials P and Q are obtained by perturbing "infinitesimally" the standard "Lotka-Volterra" case, where $P = x(a + bx + cy)$, $Q = y(e + fx + gy)$ (for convenient coefficient values, this "integrable" system has no limit cycles – the neighbouring trajectories being closed).

Adding some small second degree perturbations to these two polynomials, one can get many systems with limit cycles, but it is unknown how large the number of cycles may be.

Such bounds for the number of bifurcating limit cycles are known for the polynomial perturbations of the Hamiltonian vector fields whose Hamilton function is a polynomial of fixed degree. For the Lotka-Volterra type fields the first integral of the "integrable case"

$$H = x^{\alpha} y^{\beta} z^{\gamma},$$

where $z = 1 - x - y$ in a suitable coordinate system, is not a polynomial. The results of the infinitesimal perturbations of the polynomial Hamiltonian case are not applicable to the perturbations of the integrable Lotka-Volterra systems.

Even the number of new-born limit cycles —bifurcating from the closed orbits of the integrable system of Lotka-Volterra type, under its quadratic perturbation— is not bounded yet: One needs to bound here the number of real roots of a non Abelian integral $I(c)$ along the (non algebraic) ovals of the family $H = c$, for the above first integral H.

The 16-th Hilbert problem is one of the most important unsolved problems of mathematics.

All methods of modern mathematics provide no contribution to the problem of the limit cycles. In the long list of the Hilbert problems (which he intend to be the testament of the 19th to the 20th century) only two problems (16th and 13th) were related to topology – the domain whose development was the main contribution of the 20th century to mathematics.

To close this chapter we return to the contributions of the 18th century.

Let the variables x_k $(1 \leq k \leq n)$ have the *"quasihomogeneous degrees"* w_k, called also *weights* or *dimensions*. The *quasihomogeneous degree* of a monomial $x^u = x_1^{u_1} \cdot \ldots \cdot x_n^{u_n}$ is defined as the sum of the degrees of the variables $x_1, \ldots, x_n$, multiplied respectively by their weights: $\deg x^u = \sum w_k u_k$.

PROBLEM. Find the number of the monomials x^u of quasihomogeneous degree A.

ANSWER. Consider the power series

$$p(t) = \prod_{k=1}^{n} \frac{1}{1 - t^{w_k}} = 1 + p_1 t + p_2 t^2 + \ldots \ .$$

The number of monomials of degree A is the coefficient p_A of the term t^A.

Example. For the ordinary homogeneous degree all weights equal 1: $w_k = 1$, for $k = 1, \ldots, n$. So we get $p(t) = (1-t)^{-n}$, and hence there are $p_A = \binom{A+n-1}{A}$ homogeneous monomials of degree A.

SOLUTION. In the case of one variable, $n = 1$, the series is $p(t) = 1 + t^w + t^{2w} + \ldots$ and for each degree mw, multiple of w, there is one monomial x^m.

For n variables, one takes the product of n such "1-dimensional" series

$$\begin{aligned} p(t) &= (1 + t^{w_1} + t^{2w_1} + \ldots)(1 + t^{w_2} + t^{2w_2} + \ldots) \cdots (1 + t^{w_n} + t^{2w_n} + \ldots) \\ &= 1 + p_1 t + p_2 t^2 + \ldots \ . \end{aligned}$$

This product contains exactly one product of the factors $t^{u_1 w_1}$, $t^{u_2 w_2}$,...,$t^{u_n w_n}$ for each monomial $x^u = x_1^{u_1} \cdot x_2^{u_2} \cdot \ldots \cdot x_n^{u_n}$.

The degree in t of the product of these factors is equal to the quasihomogeneous degree of the monomial x^u.

In this way Euler invented the theory of graded algebras – "graded" means decomposed into the direct sum of homogeneous parts.

Frobenius numbers. Observe that not all integral degrees are present in the series $p(t)$:

If $d := \gcd(w_1, \ldots, w_n) \neq 1$, every degree non-multiple of d is not in the series.

If $\gcd(w_1, \ldots, w_n) = 1$, there are some gaps, but it is not difficult to show that for some $K = K(w_1, \ldots, w_n)$ every integral degree $\ell \geqslant K$ appears in the series $p(t)$.

EXERCISE. Check (and prove) that for $n = 2$ with weights $w_1 = 3$, $w_2 = 7$ every degree $\ell \geqslant 12$ is present in the series

$$p(t) = (1 + t^3 + t^{2\cdot 3} + \ldots)(1 + t^7 + t^{2\cdot 7} + \ldots) = 1 + t^3 + t^6 + t^7 + t^9 + t^{10} + \ldots \ .$$

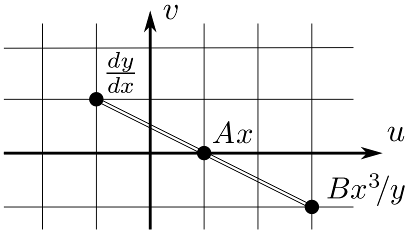
In fact, the degrees present in the series $p(t)$ form the additive semi-group generated by the weights: Its elements are all linear combinations $p_1 w_1 + \ldots + p_n w_n$ with non-negative integral coefficients $p_1, \ldots, p_n$. The value of the constant $K(w_1, \ldots, w_n)$ is the *Frobenius number* associated to $w_1, \ldots, w_n$, discussed in Section 4.4 (p. 129).

In the theory of differential equations, it is important to define the degree:

$$\deg\left(\frac{\partial}{\partial x_k}\right) = -\deg x_k.$$

For instance, the equation

$$\frac{dy}{dx} = Ax + B\frac{x^3}{y}$$



$$\frac{dy}{dx} = Ax + \frac{Bx^3}{y}$$

Figure 14.6: The Newton polygon of a differential equation helps to solve it (separating the variables by the quasihomogeneous substitution $y = zx^2$, suggested by the slope of the Newton polygon line).

is quasihomogeneous, as shows the Newton polygon of Fig. 14.6.

The "dimension theory" of physicists consists, in most cases, of applications of the Newton polyhedra to the graded algebras.

A present-day version of Newton polyhedra theory was elaborated due to some applications of the asymptotic theory of partial differential equations to biological problems.

This new theory is called "Gröbner basis theory" and it provides the most powerful computerised help to several difficult problems of algebraic geometry and of algebra.

The predecessor of this theory was the Hilbert "theological" theorem, which says that *each ideal of the ring of polynomials* (or of some other similar rings) *has a finite basis.* This result was proclaimed to be "theological" by the main experts on the theory of invariants, including Gordan, since the calculation of that finite basis is a difficult problem, which Hilbert was unable to solve, while the people working on invariant theory had earlier discovered a lot of important lists of practically useful bases.

The Gröbner basis theory suggested some powerful algorithms of calculation of those bases, which are based on the filtrations in the polynomial rings obtained from the Newton polyhedra. Gröbner had invented those tricks in the twenties of the 20th century, but his methods were unpopular until they were computerised: Those calculations are practically unachievable, unless one has a good computer. Hence the great discoveries of Gröbner remained unknown and unused for many years until the computerised algebra made them practically useful and powerful.

The present day situation is well discussed in [61].

Chapter 15

Singularities of smooth mappings

In the mathematical description of the world, discrete phenomena are perceived first, but the continuous ones have a simpler description in terms of traditional calculus. In this delicate interplay between discrete and continuous objects, Singularity Theory describes and predicts the birth of discrete objects from smooth continuous sources.

Although the diversity of possible phenomena and objects is enormous, the main lesson of singularity theory is that, in most cases, only some standard phenomena occur (the more complicated phenomena being combinations of the standard ones). So it is useful to know and to study these standard phenomena once for all times.

We shall start by showing a small set of singularities that occur in very different problems and which are as fundamental as, say, the ellipses, the hyperbolas and the parabolas. Their occurrence in very different theories is as universal as the occurrence of quadratic forms in all branches of mathematics and physics.

The reader can use and understand the results of singularity theory independently of their proofs, which often are rather technical.

15.1 Equivalences of maps and stability

One of the main objects of geometry is the study of the differential maps $f : M^m \to N^n$ of smooth manifolds. According to the problem or the property one has to study, it is often useful to introduce a suitable notion of equivalence between maps.

Example (Critical points). Consider a smooth function (Fig. 15.1).

A point where the derivative vanishes is called *critical*, and the value of the function at such a point is called the *critical value*.

A critical point is said to be *non degenerate* if the second differential at this point is a nondegenerate quadratic form. For functions of one variable,

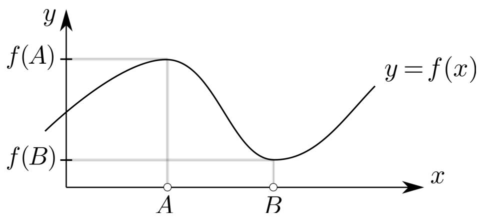


Figure 15.1: Graph of a smooth function with critical points A, B and critical values $f(A)$, $f(B)$.

this just means that $f'' \neq 0$.

Non degenerate critical points of functions of one variable are local minima points, where $f'' > 0$ (like the point B of Fig. 15.1), and local maxima points where $f'' < 0$ (like the point A of Fig. 15.1).

Consider now a slightly perturbed function $\widetilde{f}$, like $\widetilde{f} = f + \varepsilon g$, where ε is small and g is a smooth function.

The critical points of f are, as a rule, no longer critical for $\widetilde{f}$. However, if a critical point is non degenerate, the perturbed function $\widetilde{f}$ has a "perturbed" critical point in a small neighbourhood of the original critical point of the function f. For the degenerate critical points the situation may be different (Fig. 15.2).

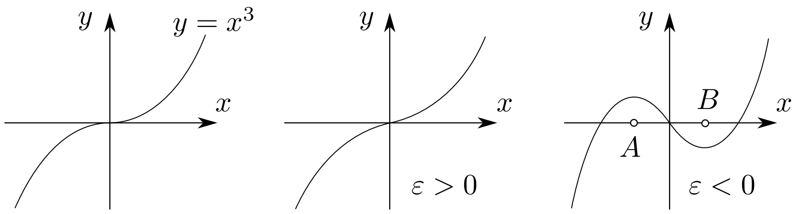


Figure 15.2: Perturbation $\widetilde{f} = x^3 + \varepsilon x$ of function $f(x) = x^3$.

Namely, the critical point $x = 0$ of the function $f(x) = x^3$ disappears* under the perturbation $\widetilde{f}(x) = x^3 + \varepsilon x$ if ε is positive, while for $\varepsilon < 0$ it bifurcates into two non degenerate critical points A and B, both close to the unperturbed critical point $x = 0$ of the unperturbed function.

*The complex critical point $x = 0 \in \mathbb{C}$ bifurcates for $\varepsilon > 0$ into two neighbouring complex conjugated non degenerate critical points of $\widetilde{f}$, but we discuss here the real manifold $M = \mathbb{R}$ for which the critical point disappears.

In this sense one says that the non degenerate critical points are stable, while the degenerate ones are not (the formal definition will be given below).

Definition. Two smooth maps $f : M^m \to N^n$ and $\widetilde{f} : \widetilde{M}^m \to \widetilde{N}^n$ are called *equivalent* if there exists a pair of diffeomorphisms

$$h : M^m \to \widetilde{M}^m \ , \quad k : N^n \to \widetilde{N}^n \ ,$$

that transform one into the other and make commutative the diagram

$$\begin{array}{ccc} M^m & \xrightarrow{f} & N^n \\ \big\downarrow h & & \big\downarrow k \\ \widetilde{M}^m & \xrightarrow{\widetilde{f}} & \widetilde{N}^n \end{array}$$

The commutativity means that $\widetilde{f} \circ h = k \circ f$, which is the abbreviated notation for the identity

$$\widetilde{f}(\widetilde{x}) = k\big(f(h^{-1}(\widetilde{x}))\big) \,, \tag{1}$$

for any $\widetilde{x} \in \widetilde{M}^m$.

In other terms, the map f is transformed to $\widetilde{f}$ by the change h of the independent variables and the change k of the dependent variables.

Problem. Are the maps $\widetilde{f}$ of Fig. 15.2, corresponding to different values of parameter ε, equivalent to each other ?

Answer. All the maps $\widetilde{f}$ for which $\varepsilon > 0$ are equivalent between them. All the maps $\widetilde{f}$ for which $\varepsilon < 0$ are also equivalent between them. The degenerate map f, which corresponds to $\varepsilon = 0$, is not equivalent to any other in the family $\{\widetilde{f}\}$.

Remark. Replacing diffeomorphisms h and k by homeomorphisms one obtains the *topological equivalence* notion.

Restricting oneself to the diffeomorphisms of the independent variables manifold M^m (taking k =identity map) one obtains the so-called *right equivalence* ("R-equivalence"): $\widetilde{f} = f \circ h^{-1}$.

The strange name "right equivalence" is due to the presence of h in the *rightmost part* of the defining identity (1).

The *left equivalence* ("L-equivalence") is defined by the changes k of the dependent variables only (h =identity): $\widetilde{f} = k \circ f$.

Of course, the initial equivalence is also called *left-right equivalence* ("LR-equivalence"). The adjectives "smooth, topological, analytic, holomorphic, formal" are sometimes added, to distinguish the cases of smooth diffeomorphisms, of homeomorphisms, of analytic or holomorphic diffeomorphisms and of formal Taylor series.

We shall mostly speak on the LR smooth equivalence, calling it simply "equivalence", and adding the epithets in other cases. [The required smoothness degrees are different for different theorems. But an admissible choice is to understand everywhere the word "smooth" in the C^∞ sense].

Returning to Fig. 15.2, we conclude that some of the (arbitrary small) perturbations $\widetilde{f}$ of the function $f(x) = x^3$ are not equivalent to it. But it is clear that any sufficiently small perturbation $\widetilde{f}$ of the function $f(x) = x^2$ is equivalent to f. In this sense we shall say that

$$f(x) = x^2 \textit{ is stable and } f(x) = x^3 \textit{ is unstable.}$$

Definition. A smooth map $f : M^m \to N^n$ is *stable* if any slightly perturbed map $\widetilde{f} : M^m \to N^n$ is equivalent to the unperturbed map. The conjugating diffeomorphisms $h : M^m \to M^m$ and $k : N^n \to N^n$ in the relation (1) are arbitrarily close to the identity maps, provided that the perturbed map $\widetilde{f}$ is sufficiently close to the initial map f.

To simplify the technical description of the words "slightly perturbed" and "sufficiently close", we can suppose M^m and N^n to be compact manifolds, measuring the differences of the maps by the C^r-metrics, taking into account the r-th partial derivatives (in some finite set of fixed local coordinate systems). One should also mention r in the definition, saying "C^r-stable", but we shall simplify the terminology, abbreviating the sentence «there exists an r_0 such that f is C^r-RL-stable for any $r \geqslant r_0$» to «f is stable».

We leave to the reader the pleasure to invent the details of the non compact case version, as well as that of the local version, which is needed to define stable critical points.

Problem. Is the smooth function of Fig. 15.1 stable ?

Problem. Suppose that all the critical points of a smooth function on the circle $\mathbb{S}^1$ are non degenerate. Is this function stable?

Answer. A counterexample is shown in Fig. 15.3: Some perturbed functions $\widetilde{f}$ have more critical values and, hence, they are topologically non-equivalent to the initial function f, which is topologically RL-unstable, say, for $f = \cos(2x)$, $x \in \mathbb{R} \pmod{2\pi}$).

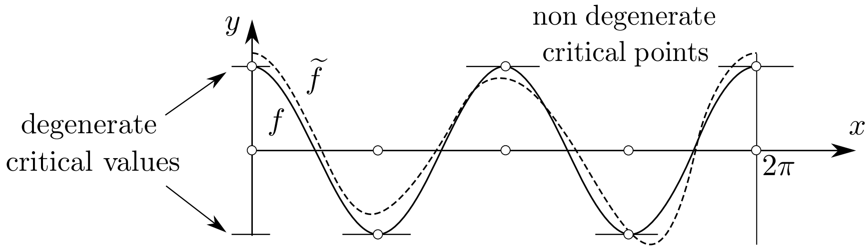


Figure 15.3: Topological unstability of a function whose four critical points are non degenerate.

Problem. Describe the topologically stable functions $f : \mathbb{R} \to \mathbb{R}$, equal to x^{n+1} for sufficiently large $|x|$ and having n critical points, when $n + 1 \leqslant 6$. The diffeomorphisms h and k are supposed here to preserve the orientations of the axes in (1).

Answer. See Fig. 15.3. The numbers $K(n)$ of topologically (or smoothly) non-equivalent functions in this problem are

$n+1$	1	2	3	4	5	6	7	8	9	10
$K(n)$	1	1	1	2	5	16	61	272	1385	6571

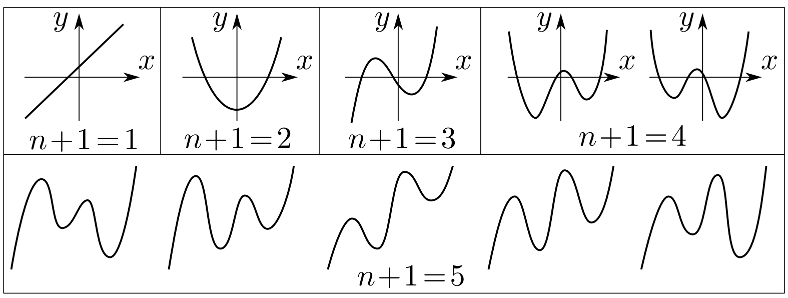


Figure 15.4: Topological RL-classification of stable polynomials $x^{n+1} + \dots$ of degree $n + 1$, having n real critical points.

Problem. Calculate the function defined by the power series

$$F(t) = \sum_{n=1}^{\infty} K(n) \, \frac{t^n}{n!} \ .$$

Answer.

$$\left(1 + \frac{t^2}{2} + \frac{5t^4}{24} + \dots\right) + \left(t + \frac{t^3}{3} + \frac{16t^5}{120} + \dots\right) = \sec t + \operatorname{tg}\, t \ .$$

Hint. Pushing to ∞ one of the critical values, decompose the graph with n critical points into the union of two graphs with k and ℓ critical points, $k + \ell = n - 1$, and deduce from such decompositions the differential equation

$$2 \, \frac{dF}{dt} = F^2 + 1 \ .$$

and, then, compare the Taylor series of the left and of the right hand side.

15.2 Morse Lemma and homotopy method

To study the general theory of smooth maps $f : M^m \to N^n$, we start from the case of one function ($n = 1$) of m variables, considering its critical points.

Definition. A critical point of a smooth function of m variables is called *non degenerate*, if the second differential of the function at that critical point is a non degenerate quadratic form.

The evident stability of non degenerate critical points of smooth functions of one variable has the following higher-dimensional sister.

Morse Lemma. *Let $O \in \mathbb{R}^m$ be a non degenerate critical point of a smooth function f. Then there exists a smooth coordinate system $(x_1, \dots, x_m)$ in some neighbourhood of O, such that the function is expressed in these coordinates in the form of its Taylor series of degree 2:*

$$f(x) = c + \sum_{k=1}^{n} a_k \ x_k^2 \ ,$$

where $a_k \in \{+1, -1\}$.

Remark. The Morse Lemma explains the occurrence of quadratic forms (and hence of ellipses, hyperbolas and so on) in most problems of geometry, calculus and physics: they are the normal forms of *arbitrary* generic functions in the vicinity of their critical points.

The *raison d'être* of algebraic geometry is similar: polynomials are either local approximations or the local normal forms of arbitrary functions or maps.

Proof of Morse Lemma. Its proof is even more important than its statement, which implies, however, the stability of f in a neighbourhood of O, even with respect to the special RL-equivalences where the dependent variable diffeomorphisms have the special form $y \mapsto y + \text{const}$.

The method of the proof is called the *homotopy method*, and the idea behind it is that it is easier to prove a *more general* statement.

Decompose the function f into its quadratic part f_2 and the remaining part f_3 of order $|x|^3$:

$$f = f_2 + f_3 \ .$$

Connect f to its quadratic part by the homotopy

$$F(x,t) = f_2(x) + t\ f_3(x) \ , \qquad 0 \leqslant t \leqslant 1 \ .$$

Denote the members of this connecting family of functions by F_t,

$$F_t(x) = F(x,t) \ .$$

We wish to prove the equivalence of the function $F_1 = f$ to $F_0 = f_2$, that is, to find a local diffeomorphism $h : (\mathbb{R}^m, O) \to (\mathbb{R}^m, O)$ such that

$$F_1\bigl(h(x)\bigr) = F_0(x)$$

in some neighbourhood of the point $x = O$.

The homotopy idea is to try to prove the equivalence to F_0 of *all* the connecting functions F_t $(0 \leqslant t \leqslant 1)$, that is, to find a family of diffeomorphisms h_t verifying the identity

$$F_t\bigl(h_t(x)\bigr) = F_0(x) \ , \tag{2}$$

that is, $F_t \circ h_t = F_0$, for all $0 \leqslant t \leqslant 1$, in some neighbourhood of the point $x = O$.

Now to find the family $\{h_t\}$ of the conjugating diffeomorphisms that verify identity (2), we construct them step by step, studying the needed velocity of deformation of h_t with respect to t. This velocity vector-field v_t, which depends on t, is defined as

$$\frac{dh_t(x)}{dt} = v_t\big(h_t(x)\big) \ .$$

The identity (2) implies an "infinitesimal commutativity condition" for the field v_t, called the "*homological equation*", that we shall write and then solve in the unknown fields v_t.

We can say that we decompose the commutative diagram of page 535 into thin horizontal slices conjugating $F_{t+\varepsilon}$ to F_t. Then we replace the thin commutativity condition of a slice by its infinitesimal version (which is the homological equation for the vector field that represents the small vertical side of the slice), and then we integrate the resulting infinitely thin horizontal diagrams to get the big diagram that transforms F_1 to F_0 – Fig. 15.5.

$$\begin{tikzcd}
M^n \arrow[r, "F_1"] \arrow[d] & \mathbb{R} \arrow[d, "\mathrm{id}"] \\
M^n \arrow[r, "F_{t+\varepsilon}"] \arrow[d, "\widetilde{x} \mapsto \widetilde{x} + \varepsilon v_t(\widetilde{x})"'] & \mathbb{R} \arrow[d, "\mathrm{id}"] \\
M^n \arrow[r, "F_t"] \arrow[d] & \mathbb{R} \arrow[d, "\mathrm{id}"] \\
M^n \arrow[r, "F_0"] & \mathbb{R}
\end{tikzcd}$$

Figure 15.5: Integration of the infinitely thin commutative diagrams of the homotopy method.

Remark. Consider the action $A : G \times V \to V$ of a Lie group G on a smooth manifold V. The manifold V is subdivided into the orbits of the action A. The orbit of a point $p \in V$ is the submanifold $A(G \times \{p\}) \subset V$.

The tangent space at p to the orbit of the point p, denoted by $W(p)$, is the image of the partial derivative map

$$A_{*e} : (T_eG \cdot p) \to T_pV \ , \quad W(p) = (A_{*e}\mathfrak{g}) \subset T_pV \ .$$

The homotopy method connects two points x and y of the same orbit by a smooth curve γ whose tangent vector at each point z of it belongs to the tangent space $W(z)$ of the orbit of z. However, to find just such a connecting curve is insufficient (the curve of the following theorem shows it).

Theorem 1. *There exists a smooth curve γ that connects two points p and q of different orbits, whose tangent vector at each point z of it belongs to the tangent space $W(z)$ of the orbit of the point z.*

Proof. Let V be the complex projective line $\mathbb{C}P^1$ provided with affine coordinate $z = x + iy$ and let G be the group $\mathrm{SL}(2, \mathbb{R})$. Consider its natural action which, in the affine coordinate, is defined by the standard formula.

$$A\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}, z\right) = \frac{az + b}{cz + d} \ .$$

This action has three orbits: The real line $\mathbb{R}\mathrm{P}^1 \subset \mathbb{C}P^1$, on which $y = 0$, and the two half-planes $y > 0$ and $y < 0$.

The point p where $z = 0$ and the point q where $z = 1 + i$ belong to different orbits. The smooth parabola $\gamma(t) = (t, it^2)$ connects the points p and q, and its tangent vector at every point z of it belongs to the tangent space $W(z)$ of the orbit of the point z. □

The homotopy method construction is different because it provides a vector-field of tangent vectors along the curve γ, which is the image of a 1-parameter set of vectors $v_t \in T_e G$ that depends smoothly on t.

Technically speaking, we differentiate the commutative diagram identity (2) with respect to the parameter t. The resulting infinitesimal commutativity condition is the *homological equation*

$$\frac{\partial F_t}{\partial t} + F_{t*}\, v_t = 0 \ . \tag{3}$$

We have not written explicitly the space arguments of these functions: They are $\widetilde{x} = h_t(x) \in \mathbb{R}^m$. The space derivative F_{t*} of the function F_t is calculated here with respect to the space argument $\widetilde{x} = h_t(x)$ of the function F_t, rather than with respect to the initial condition x on which this argument $\widetilde{x}$ depends.

The homological equation (3) should be satisfied at all points $\widetilde{x}$ for every t between 0 and 1 and, hence, we have the right to denote these points $\widetilde{x}$ by any symbol, including even x (avoiding, however, to mix it with the initial point x).

We have to solve the homological equation (3), which is a *linear algebraic* (non differential) *equation with respect to the unknown vector-field v_t.*

Note that the function

$$F_* \ v = \sum_{k=1}^{m} \frac{\partial F}{\partial x_k} \ v_k \tag{4}$$

is the velocity of change of F produced by the action of the infinitesimal diffeomorphisms whose velocity of change is described by the vector-field v.

The combinations (4) of the partial derivatives of F, with arbitrary coefficients v_k, form a vector-space and even an ideal, in the space of functions, called the *gradient ideal of the function* F. In the language of p. 540, this gradient ideal is the *tangent space* $W(F)$ *at* F *of the orbit of the point* F under the action of the right-equivalence group G.

Therefore, the homological equation (3) means that the homotopy velocity vector $\partial F_t/\partial t$ (called "tangent vector to the curve γ" on page 540) belongs to the tangent space W of the orbit at the corresponding point F_t of the "curve $\gamma = \{F_t : 0 \leqslant t \leqslant 1\}$" at any moment t.

We shall prove below that the homotopy velocity vector genuinely belongs to the tangent space of the orbit. We shall also construct a solution v_t of the homological equations (3) that is valid for all the time moments $0 \leqslant t \leqslant 1$ of the homotopy and that depends smoothly on the time moment t.

It is the smoothness in t of the solution of the homological equation what is absent in the paradoxical curve (of page 540) that joins two different orbits. Try to find a smooth solution there!

In our particular problem, $\partial F_t/\partial t = f_3$ is the velocity of the perturbation and we have to find the velocity of the diffeomorphisms changing v_t that would compensate the perturbations.

Let us investigate the linear operator F_{t*} that acts on the field v_t in the homological equation (3). To simplify the formulae, we can choose from the very beginning the special system of coordinates that reduces the second differential of f at 0 to the simple normal form:

$$f_2 = c + \sum_{k=1}^{m} \varepsilon_k \ x_k^2 \ .$$

To simplify even more, we shall suppose the second differential to be positive definite (all $\varepsilon_k = +1$), leaving to the reader to rewrite the next calculations for the non degenerate quadratic forms of arbitrary signature.

Thus, $f_2 = c + \sum_{k=1}^{m} x_k^2$, grad $f_2 = 2x$ and, hence, the linear operator f_{2*} sends the unknown field $v = \sum v_k \ \partial/\partial x_k$ to the function $\sum_{k=1}^{m} 2x_k v_k$, which ought to be $-f_3$.

The Taylor series of the function f_3 starts from the 3th order terms, therefore *it can be represented in the form*

$$f_3 = \sum_{k=1}^{m} x_k \; u_k(x) \; , \tag{5}$$

where the Taylor series of the functions $u_k(x)$ start at least from quadratic terms.

This decomposition, called "Hadamard Lemma", is evident for Taylor series or for holomorphic functions.

To prove Hadamard Lemma for smooth functions, it suffices to divide by x a smooth function f of one variable, vanishing at the origin. The ratio is smooth at the origin too.

Proof.

$$f(x) = \int_0^1 \left(\frac{d}{dt} f(tx)\right) \; dt = \int_0^1 x \; g(tx) \; dt = x \int_0^1 g(tx) \; dt \; ,$$

where $g = f'$ is a smooth function. Hadamard lemma is proved for the functions of one variable.

For a function f of two variables, vanishing at $(0,0)$, one writes

$$f(x_1, x_2) = f(x_1, 0) + g(x_1, x_2),$$

where $g = 0$ for $x_2 = 0$. Then one decomposes the summands according to the case of one variable:

$$f(x_1, 0) = x_1 \; u_1(x_1) \; , \quad g(x_1, x_2) = x_2 \; u_2(x_1, x_2) \; ,$$

considering x_1 as a parameter for the second case.

Hadamard Lemma is thus proved for $f(x_1, x_2)$.

Similarly, the case of m variables is reducible to that of $m-1$ variables. $\square$

We conclude that the homological equation (3) is solvable for $t = 0$:

$$-f_3 = F_* v$$

for the field v with components $v_k = -u_k/2$, where the functions u_k are defined by the decomposition (5) of Hadamard Lemma.

The same argument solves the homological equation (3) for every value of t. Indeed, consider the partial derivatives

$$\frac{\partial F_t}{\partial x_k} = y_k \ , \quad k = 1, \dots, m \ .$$

These m functions in $\mathbb{R}^m$ have the same linear parts of the Taylor series $2x_k$ at the point 0, as for $t = 0$, since the perturbing term tf_3 is of order $|x|^3$. Therefore, these m functions form a smooth coordinate system in a neighbourhood of $O \in \mathbb{R}^m$, which depends smoothly on the parameter t.

Writing the homological equation (3) in this coordinate system, we get the equation

$$-f_3 = \sum_{k=1}^{m} y_k \ v_k \ , \quad \text{where} \quad w = \sum_{k=1}^{m} v_k y \ \frac{\partial}{\partial y_k} \ ,$$

which by the Hadamard Lemma is solvable in a neighbourhood of $O \in \mathbb{R}^m$ for $0 \leqslant t \leqslant 1$. The resulting functions v_k vanish at O.

Thus the homological equation (3) is solved, and the field v_k vanishes at O for all t.

Now to find the required diffeomorphisms h_t, it suffices to solve the ordinary differential equation

$$\frac{dh_t}{dt} = v_t \tag{6}$$

defined by the time-dependent smooth vector field v_t in a neighbourhood of $O \in \mathbb{R}^m$, with the initial conditions $h_0(X) = X$, to obtain the value $h_t(X)$ of the diffeomorphism h_t at the point $X \in \mathbb{R}^m$.

The theory of ordinary differential equations provides solutions for a short time interval $|t| < \delta$. We need the diffeomorphism h_1 to reduce the function F_1 to the normal form F_0. Hence, we need to investigate the prolongation of the solutions of the differential equation (6) to larger values of the time t.

Fortunately, we know that $v_t(0) = 0$ for any t and, hence, we know the solution $h_t \equiv 0$ for the initial condition $X = O \in \mathbb{R}^m$, for an arbitrary time interval.

The theory of ordinary differential equations states that for any fixed time interval, say, for $0 \leqslant t \leqslant 1$, there exists a neighbourhood of the initial condition for which the solution exists along this time interval, such that for all the initial points X of the neighbourhood there exist also solutions and these solutions depend smoothly on the initial conditions.

In consequence, we obtain the diffeomorphisms h_t from the system (6).

These diffeomorphisms verify the identity (2) because the left hand part is equal to the right hand part for $t = 0$ and the derivative in t of the left hand part is 0 for all t in the interval $0 \leqslant t \leqslant 1$, according to the homological equation (3), of which the field v_t is a solution.

We conclude, therefore, that all the functions F_t are equivalent to F_0, and in particular that $F_1 = f$ is equivalent to the Morse normal form $F_0 = f_2$. The Morse Lemma is proved.

Corollary. *Suppose that the rank of the second differential of the function f at a critical point is r. Then the function is reducible, by a smooth change of variables, to the form*

$$f(x,y) = c + \sum_{k=1}^{r} \varepsilon_k \, x_k^2 + g(y) \ , \tag{7}$$

where the Taylor series of the smooth function g starts at least from the cubic terms ($|g(y)| \leqslant C|y|^3$).

Here $(x_1, \ldots, x_r, y_1, \ldots, y_s)$ is a smooth coordinate system whose origin $O \in \mathbb{R}^m$, where $m = r + s$, is the critical point under consideration.

Proof. It is easy to reduce the quadratic part of the Taylor series to the above normal form by a linear change of variables. Now, considering f as a function of r arguments x, depending on s parameters y, we see that for the parameter value $y = 0$ it has a non degenerate critical point at $x = 0$.

The implicit function theorem implies that for each neighbouring value of the parameter $y \in \mathbb{R}^s$, there exists a neighbouring critical point $x = z(y)$, which depends smoothly on y near $y = 0$ (we use the non degeneracy of the quadratic part in x).

The y-dependent diffeomorphisms of the x-space,

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} X = x - z(y) \\ y \end{pmatrix} \ ,$$

transform our y-dependent family of functions of $x \in \mathbb{R}^r$ into a y-dependent family of functions of $X \in \mathbb{R}^r$, all of them having a common critical point $X = 0$. If y is sufficiently small, this critical point is non degenerate because it is non degenerate for $y = 0$ and depends on y continuously.

Applying the Morse lemma to this family of functions, we obtain a family of local diffeomorphisms of the space $\mathbb{R}^r$ near $X = 0 \in \mathbb{R}^r$, which depends smoothly on y. This family reduces all these functions of the argument X to the Morse normal form,

$$f = c(y) + \sum_{k=1}^{r} \varepsilon_k \, X_k^2 \, .$$

The linear and quadratic terms of the Taylor series of the smooth function $c = c(y)$ must be zero. Indeed, the linear terms would make the point $(y = 0, X = 0)$ non critical, and the quadratic terms would increase the rank of the second differential of f at 0. The corollary is proved. $\square$

Corank 1 Singularities: The $\mathbf{A_n}$ Series. In the case $s = 1$, it is easy to reduce the function g of (7) in the above corollary (by a diffeomorphic change of the coordinate y) to the form $g(y) = \pm y^n$, where Cy^n is the first nonzero term of the Taylor series of the function g at the origin.

Therefore if the rank of the second differential of a function of m variables at a critical point is $m - 1$ (i.e. if its corank equals 1), then the function is reducible (by a smooth change of variables) to the normal form

$$A_{n-1} \; : \; f = c \pm y^n + \sum_{k=1}^{m-1} \varepsilon_k \, x_k^2 \, ,$$

provided that the Taylor series of the function g does not vanish. These "corank one critical points" (called A_{n-1} *critical points*) are the product of the collision of $n - 1$ non-degenerate critical points and are closely related to the Lie groups $\mathrm{SL}(n, \mathbb{C})$ and $\mathrm{SU}(n, \mathbb{C})$.

Example. For functions of two variables the simplest degenerations are described by the series of corank one normal forms (see Fig. 15.6)

$$A_2 \; : \; f = y^3 \pm x^2 \; , \quad A_3 \; : \; f = \pm y^4 \pm x^2 \, , \ldots ,$$

whose general term is $A_n \; : \; f = \pm y^{n+1} \pm x^2$.

A generic smooth function $f : M^m \to \mathbb{R}$ on a manifold M^m is a stable map whose critical points are non degenerate and have different critical values.

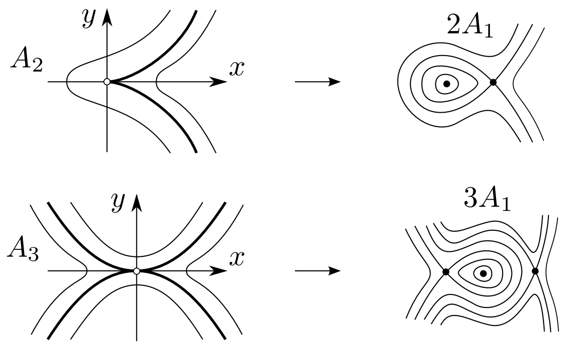


Figure 15.6: The level lines of two functions, near the simplest degenerate critical points "A_2" and "A_3", and those of their perturbations.

15.3 Singularities of maps between surfaces

For each pair of dimensions $(m; n)$ there is a theory for the classification and stability of the singularities of generic smooth maps $f : M^m \to N^n$.

In case $m = n = 2$, the generic maps between surfaces $f : M^2 \to N^2$ are locally equivalent at every point to one of the following three Whitney normal forms:

$$(1)\ \begin{cases} y_1 = x_1\ , \\ y_2 = x_2\ ; \end{cases} \qquad (2)\ \begin{cases} y_1 = x_1^2\ , \\ y_2 = x_2\ ; \end{cases} \qquad (3)\ \begin{cases} y_1 = x_1^3 + x_1 x_2\ , \\ y_2 = x_2\ . \end{cases}$$

Here (x_1, x_2) are local coordinates in M^2, (y_1, y_2) – in N^2, and the formula

$$y_1 = f_1(x_1, x_2)\ , \quad y_2 = f_2(x_1, x_2)$$

means that the map $f : M^2 \to N^2$ sends the point with coordinates (x_1, x_2) in M^2 to the point with coordinates $\big(f_1(x_1, x_2), f_2(x_1, x_2)\big)$ in N^2.

Some algebraists would write $f^* y_1 = f_1$, $f^* y_2 = f_2$ on M^2, which makes the notations less understandable.

Example. The orthogonal projection of the sphere to a plane is locally reducible to the normal forms (1) and (2) – Fig. 15.7. A point P of the sphere M^2 is *singular* for the projection $f : \mathbb{S}^2 \to \mathbb{R}^2$ if the derivative

$$f_{*P} : T_P \mathbb{S}^2 \longrightarrow T_{f(P)} \mathbb{R}^2$$

is a degenerate linear map, that is, if the Jacobian determinant vanishes at the point P. This happens only at the equator of the sphere.

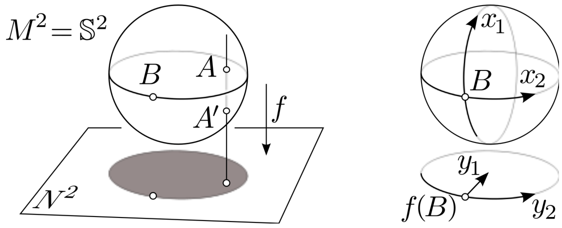


Figure 15.7: The fold singularities of a smooth map between two surfaces.

The point A on Fig. 15.7 is not singular for the projection f, which is a local diffeomorphism in its neighbourhood. In suitable local coordinates around the point A, this map can be written in the above normal form (1).

At the point B of the equator, shown in Fig. 15.7, the situation is different. The map is not even a local diffeomorphism, its image does not cover all the neighbourhood of the projected point $f(B)$, and the number of preimages of the point $f(B)$ is one, while the number of preimages is equal to two or to zero at some neighbouring points of $f(B)$ in the plane N^2.

However, it is clear that choosing the longitude x_2 as one of the coordinates on the sphere, near the point B, and using the corresponding angular coordinate y_2 on the plane, we reduce our map f to the one-parameter family of smooth maps of the meridians to the radial rays. At the point B such a map has a non degenerate critical point. Applying the Morse lemma, we reduce the map f, in some neighbourhood of B, to the normal form (2) of the above list of Whitney.

This singularity is called a *fold*. It is a stable singularity, as it is a Morse singularity of a smooth function. Looking at the (presumably smooth) faces of our friends, we see a lot of visible contours, like the profile line, which are the curves of critical values of some folds of the projections of our friends faces on the retina surface of our eyes – Fig. 15.8.

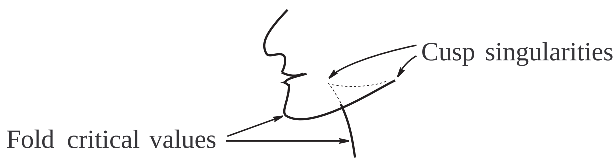


Figure 15.8: Singularities of an apparent contour

The third Whitney singularity, called "cusp singularity", is much more complicated, but it is also stable and generic. We can see it everywhere, but people did not notice it before the Whitney paper of 1955 where he proved its genericity, stability and universality.

As an example of the Whitney cusp singularity, consider the surface M^2 defined in Euclidean 3-space, with coordinates (x_1, x_2, y_1), by the equation

$$y_1 = x_1^3 + x_1 x_2 \ .$$

This smooth surface M^2 is the graph of the right hand side polynomial. Thus, x_1 and x_2 form a coordinate system on the surface M^2. The projection f (vertically down in Fig. 15.9) of the Whitney normal form (3) sends the point (x_1, x_2, y_1) of M^2 to the point $(y_1, y_2 = x_2)$ of the plane $N^2 = \mathbb{R}^2$.

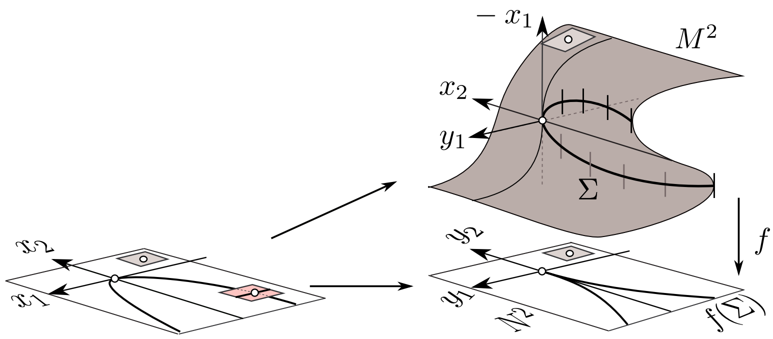


Figure 15.9: The Whitney cusp singularity of the projection of a surface.

To find the critical points and the folds of the map f, we write its Jacobian matrix in our coordinate system,

$$(f)_* = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 3x_1^2 + x_2 & x_1 \\ 0 & 1 \end{pmatrix} \ .$$

The Jacobian determinant

$$J = \det(f_*) = 3x_1^2 + x_2$$

vanishes on the parabola $x_2 = -3x_1^2$. The singular points of the projection f form a space curve Σ on surface the M^2, which may be parametrised by the coordinate x_1,

$$\Sigma : \quad x_1 = t \ , \ x_2 = -3t^2 \ , \ y_1 = -2t^3 \ .$$

The critical values of the map f form the curve $f(\Sigma)$, which is the projection of the space curve Σ to the plane N^2. It is a semicubic cusp

$$f(\Sigma): \quad y_1 = -2t^3 \ , \quad y_2 = -3t^2 \ .$$

The singularity is a fold everywhere along the curve of critical points Σ, except at the point 0. In the neighbourhood of this point the "Whitney cusp map" is even topologically non-equivalent to a fold. For instance, some neighbouring points $y \in N^2$ have 3 preimages, some others only one.

The kernel $f_{*x}^{-1}(0)$ of the degenerate derivative

$$f_{*x} : T_x M^2 \longrightarrow T_{f(x)} N^2$$

at a point x of Σ is a tangent line to M^2 at x (which is vertical at Fig. 15.9).

These vertical lines form a field of lines along the smooth singularities curve $\Sigma \subset M^2$. When point x moves along Σ, the kernels rotate in the tangent planes to M^2, and the Whitney cusp normal form corresponds to the coincidence of the kernel direction with the tangent direction of the curve of singularities Σ, at 0.

Whitney's theorem states that (i) *this cusp singularity is stable*, and (ii) *a generic smooth map $f : M^2 \to N^2$ has no other singularities.*

Therefore, at any point of M^2, a generic smooth map is locally equivalent to one of the three maps (1), (2), (3) of the above list.

For example, we see one half of the semicubic cusps of $f(\Sigma)$ *at the end-points of the visible contour lines of non transparent smooth surfaces*, like our own visible contours (Fig. 15.10 right). They are better seen on transparent surfaces (in Fig. 15.10, we show two such cusps, visible on a bottle).

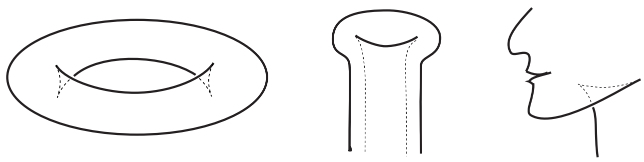


Figure 15.10: Visible semicubic Whitney cusp singularities: two on a toric tyre, two on a bottle surface and one on the projection of a face.

Some smooth maps between surfaces have other singularities than the three generic ones described above. But they are unstable: A generic small smooth perturbation of the map decompose them into the generic ones – similarly to the decomposition of the non generic singularity x^3 of Fig. 15.2 into two generic Morse singularities.

Example. The squaring of the complex numbers, $f(z) = z^2$, can be considered as a smooth map of the real plane with coordinates x and y, where $z = x+iy$, into the real plane with coordinates u and v;*, where $z^2 = w = u + iv$.

This smooth map (called the "complex folding")

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^2 - y^2 \\ 2xy \end{pmatrix}$$

is evidently different from any of the three generic ones. It is non-equivalent to them at $z = 0$ both in the diffeomorphism sense and in the homeomorphism sense. According to the Whitney theorem, it bifurcates into several generic singularities under any generic small smooth real perturbation.

Problem. Show that the real perturbation $f_\varepsilon = z^2 + 2\varepsilon\overline{z}$ suffices to decompose the singularity of the complex squaring map into generic singularities (and it is a generic perturbation), where $\overline{z} = x - iy$ and ε is a small real parameter. Then, find the critical points of the maps $f_\varepsilon : \mathbb{R}^2 \to \mathbb{R}^2$, their critical values and the kernels of the derivatives at the critical points.

Solution. The answer is shown in Fig. 15.11: From the Jacobian matrix

$$(f_{*z}) = \begin{pmatrix} 2x + 2\varepsilon & -2y \\ 2y & 2x - 2\varepsilon \end{pmatrix} ,$$

we get $J(z) = 4(x^2 + y^2 - \varepsilon^2)$. The critical set is thus the circle $|z| = \varepsilon$.

To find the kernel, write $df = 2z\, dz + 2\varepsilon d\overline{z}$ and solve the equation $df = 0$ in the variable dz. We obtain at the point $z = \varepsilon e^{i\varphi}$ of argument φ the value of the differential $df = 2\varepsilon(e^{i\varphi}\, dz + d\overline{z})$. Denoting by ϑ the argument of the complex number dz, we find that the direction $\vartheta = \vartheta(\varphi)$ of the kernel at the critical point $z = \varepsilon e^{i\varphi}$ is given by the equation

$$\varphi + \vartheta = \pi - \vartheta \pmod{\pi} ,$$

which means that $\vartheta = (\pi - \varphi)/2 \pmod{\pi}$.

This direction is tangent to the circle Σ of singularities at the points where $\vartheta = \varphi + \pi/2 \pmod{\pi}$, therefore we get $3\varphi/2 = 0 \pmod{\pi}$, $\varphi = 0$, $2\pi/3$, $4\pi/3$. Thus we get three Whitney cusp singularities, connected by three arcs of folds. The set of critical values is a hypocycloid with three cusps, also called *deltoid*,

$$w = \varepsilon^2(e^{2i\varphi} + 2e^{-i\varphi}) .$$

The resulting picture is shown in Fig. 15.11.

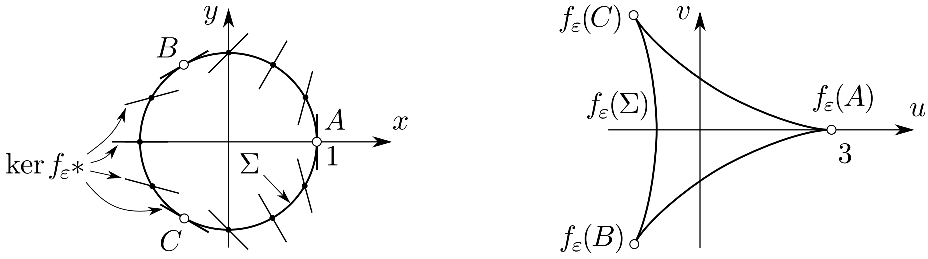


Figure 15.11: A generic real perturbation of the complex squaring map $f(z) = z^2$.

Problem. Find the generic perturbations of the so-called “handkerchief map”, $f(x, y) = (x^2, y^2)$ of the real plane to the real plane, their critical points, critical values, the kernel of the derivative at the critical points and the topological structure.

15.4 Singularities of maps from surfaces to 3-manifolds

The generic singularities of the smooth maps of surfaces into the 3-dimensional spaces or manifolds also have a sufficiently simple description.

Of course, for such a map $M^2 \to N^3$, two different smooth branches of the image surface may intersect transversely along a curve or even three branches may intersect at isolated points. But the intersection of four branches is not generic – this theorem is called “*the Gibbs rule*” in thermodynamics.

Problem. Suppose that you see a high voltage electric transmissions line that contains several threads, which are not exactly straight lines and which we shall suppose to be generic space curves.

Travelling in the fields, you try to find a place from which you see the maximal number of mutually “crossing” lines on the same ray of vision. What will be this maximal number?

Answer. The Gibbs principle answer is: two if you are lazy, three if you follow a road, four if you investigate the field, but never five if the observed curves are generic.

Problem. Similar question: What is the maximal number of tangency points of a straight line in $\mathbb{R}^3$ with a generic smooth non convex surface of a potato?

Answer. Here the answer is also four, provided that the surface of the potato is sufficiently complicated and non convex, while five tangencies are possible only for the non generic potatoes.

The intersections of different branches of the image of a smooth surface for a generic smooth map to $\mathbb{R}^3$ are not local critical points: The derivatives of the map at the corresponding points are non degenerate, the implicit function theorem is applicable, and every branch of the image surface is a smooth submanifold.

The only *generic local singularity* of smooth maps $f : M^2 \to N^3$ is the so-called *Whitney-Cayley umbrella* surface. We describe it now.

Consider in $\mathbb{R}^3$, equipped with coordinates (x, y, z), the surface defined by the equation

$$y^2 = zx^2 \ . \tag{8}$$

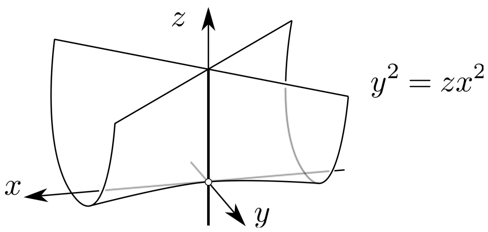


Figure 15.12: The Whitney-Cayley umbrella surface.

To understand this surface, consider its intersections with the horizontal planes, $z = \text{const}$ – Fig. 15.12. Each intersection consists of two straight lines that intersect each other at $x = y = 0$, if $z > 0$.

When z lowers to 0, the 2 lines collide, disappearing afterwards, for $z < 0$.

To see their behaviour better, observe that the surface intersects the vertical plane $x = 1$ along the parabola $z = y^2$. The line $x = y = 0$ is a regular self-intersection line of the surface in its part where $z > 0$. Its part where $z < 0$, “the handle of the umbrella”, also belongs to the algebraic surface (8). It is the real intersection line of two complex (conjugated) smooth branches, where $z < 0$, of algebraic surface (8).

Whitney discovered a generic and stable smooth map $f : \mathbb{R}^2 \to \mathbb{R}^3$, whose image is the umbrella surface without the handle. Denoting by u and v the coordinates in $\mathbb{R}^2$, we can write this Whitney normal form in the following way:

$$y = uv \ , \quad z = u^2 \ , \quad x = v \ .$$

The relation $y^2 = zx^2$ is obviously fulfilled. The proofs of the stability and the genericity of this Whitney map are too long for this book (see, for instance, [103]).

Whitney proved also that the *stable, generic maps form an open dense set in the space of smooth maps* $f : M^2 \to N^3$. It means that all different singularities of a smooth map of a surface into a 3-manifold bifurcate into some systems of Whitney umbrellas under a generic deformation of that map.

Problem. Consider a generic section of the Whitney umbrella by a smooth surface that contains the singularity O, and study its behaviour under the generic deformations of the dissecting surface, no longer containing O. A typical example is provided by the family of surfaces $x + z = t$, depending on a parameter t.

Answer. It is the $\gamma \to U$ metamorphosis, Fig. 15.13, which has a separating semicubic cusp at the critical moment ($t = 0$ in the preceding example).

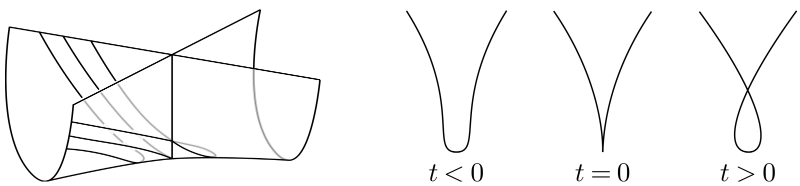


Figure 15.13: The typical $\gamma \to U$ evolution of the sections of an umbrella.

The Whitney–Cayley umbrella had been studied by many scientists centuries earlier. A. Cayley called it “umbrella” in spite of the bad protection from the rain, but appreciating the convenient handle.

In many works on algebraic geometry this singularity appeared under the name “vertex”, especially in the attempts to extend Plücker theory of singular points of algebraic curves to higher-dimensional surfaces.

But only Whitney observed the universality of this singularity that occurs in generic maps from smooth surfaces to a 3-space, independently of any algebraic geometry.

15.5 The swallowtail surface

For generic maps from 3-manifolds to 3-manifolds ($m = n = 3$), the most complicated generic singularity is a direct generalisation, to dimension 3, of the 3rd Whitney (cusp) singularity of maps from surfaces to surfaces.

The Whitney singularity of a smooth map $f : \mathbb{R}^3 \to \mathbb{R}^3$ is reducible to the following normal form (it is useful to guess the $\mathbb{R}^n \to \mathbb{R}^n$ version):

$$f\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} y_1 = & x_1^4 + x_1^2 x_2 + x_1 x_3 \\ y_2 = & x_2 \\ y_3 = & x_3 \end{pmatrix} .$$

We shall not repeat the calculations of the Jacobian determinant J of the surface of the singularities, $\Sigma : J(x) = 0$, and of its line Σ' where the kernels of f_* are tangent to Σ. All these submanifolds are smooth, and the most singular point $\Sigma'' = O$ of the line Σ' is distinguished by the tangency of the kernel to the curve Σ'.

The set of critical values, $f(\Sigma)$, is a singular surface in $\mathbb{R}^3$, which is a natural generalisation of the semicubic cusp in the plane (formed by the critical values of the 2-dimensional Whitney cusp singularity).

There exist about a hundred of independent definitions of this remarkable surface $f(\Sigma)$, called the *swallowtail*, and therefore about 10^4 theorems that claim that all these definitions lead to diffeomorphic varieties.

As we have seen in Sec. 5.1.1 of Chapter 5, one of the earliest appearances of this variety was in the Kronecker studies of real polynomials:

Swallowtail Surface. The *swallowtail surface* is the set of those values of the real parameters (a, b, c) for which the degree 4 polynomial

$$x^4 + ax^2 + bx + c = (x-u)^2\ (x^2 + 2ux + v) \tag{9}$$

has a real multiple root $(x = u)$.
[The *complex swallowtail* is defined similarly.]

Sometimes the values of the real parameters that correspond to complex multiple roots, are included in the real swallowtail, to make it the algebraic subvariety $\Delta(a, b, c) = 0$ of the space $\mathbb{R}^3$, where Δ is the discriminant.

To understand the geometry of the swallowtail surface, parametrise it by the parameters u and v, calculating from (9):

$$a = v - 3u^2\ ,\ b = 2u^3 - 2uv\ ,\ c = u^2 v\ .$$

Now to study the plane sections $a =$ const of the parametrised surface, we substitute $v = a + 3u^2$ in the expressions of b and c. The resulting formulas provide the parametrisation of the plane section curves

$$b = -4u^3 - 2au\ ,\quad c = 3u^4 + au^2\ ,$$

where u is the parameter of the curve and a indicates the corresponding section $a = \text{const}$. To draw these parametrised curves one starts by drawing the graphs of $b(u)$ and $c(u)$ for different values of a. We show the final picture in Fig. 15.14, obtained from the union of the plane sections.

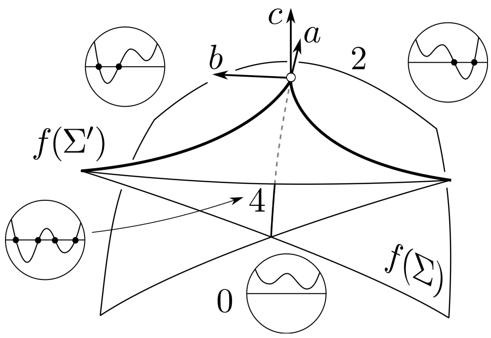


Figure 15.14: The swallowtail surface formed by the polynomials $x^4 + ax^2 + bx + c$ of zero discriminant.

The complementary space, formed by the polynomials (9) with no real multiple roots, is subdivided by the swallowtail surface into three connected parts (Fig. 15.14). These parts are formed by the polynomials with 0, 2 or 4 real roots.

Another description of the swallowtail surface says that *it is formed by the tangent straight lines of the space curve*

$$A = t^2 \ , \quad B = t^3 \ , \quad C = t^4 \ .$$

This curve is the cuspidal edge of the swallowtail surface ($f(\Sigma')$ in Fig. 15.14).

15.6 Maps to higher-dimensional spaces

Among other pairs of dimensions we shall now consider the pair $(1, n)$, which corresponds to the space curves $f : M^1 \to N^n$, where $M^1 = \mathbb{R}$ or $\mathbb{S}^1$.

The generic plane curves have no local singularities because the vanishing of the derivative vector $f_* = (\, df_1/dx,\, df_2/dx \,)$ requires *two* independent conditions on *one* unknown x.

Thus, the only singularities of the image curve are the transverse self-intersections of different branches, like for the "∞" curve.

A smooth map is called *immersion* if its derivative map is injective at every point, that is, if the kernel of its derivative map is zero at every point.

Example. The generic maps of a smooth curve to a surface are immersions. In general, the image of an immersion $M^m \to N^n$ can have self-intersections, but every branch is locally a smooth submanifold of dimension m.

Generic maps $f : M^1 \to N^{n \geqslant 3}$ are embeddings, since the self-intersections are easily removed by a small generic perturbation. The higher dimensional version of this fact is the following embedding theorem.

Theorem 2. *The generic maps of an m-dimensional manifold M^m into an n-dimensional manifold N^n are embeddings if n is sufficiently large, namely, if $n \geqslant 2m + 1$.*

This embedding theorem follows from the Gibbs principle proved on p. 21 because we can replace maps to $\mathbb{R}^{2m+1}$ by maps to any $2m + 1$-dimensional smooth manifold (using its local diffeomorphism to $\mathbb{R}^{2m+1}$).

Sard's Lemma enables us to say even more :

Theorem 3. *Any smooth submanifold of a Euclidean space $\mathbb{R}^N$ can be embedded into $\mathbb{R}^{2m+1}$, and this embedding can be chosen in an arbitrary small neighbourhood of any map of M^m to $\mathbb{R}^{2m+1}$.*

With the same reasonings, one proves also the following topological result for immersions: *Any smooth submanifold of any Euclidean space* $\mathbb{R}^N$ (and hence every existing smooth manifold) *admits an immersion to* $\mathbb{R}^{2m}$ (and, hence, to any $2m$-dimensional manifold).

15.7 Simple singularities and bad dimensions

To have a general overview of (the singularities of) the generic smooth maps $f : M^m \to N^n$, let us draw the table of all pairs of dimensions (Fig. 15.15) which includes the preceding cases and more. Let us explain the notations.

The symbol "⬭" means the absence of any singularity : the generic maps embed M^m into N^n. The symbol "∞" means the absence of local singularities : the generic maps are immersions whose singularities, if any, are at most intersections of different branches of the image. The A_k series (corank 1 singularities) has been seen above.

The symbol "☹" represents the so-called *bad dimensions* where a discrete classification of generic maps is impossible due to the presence of "moduli"

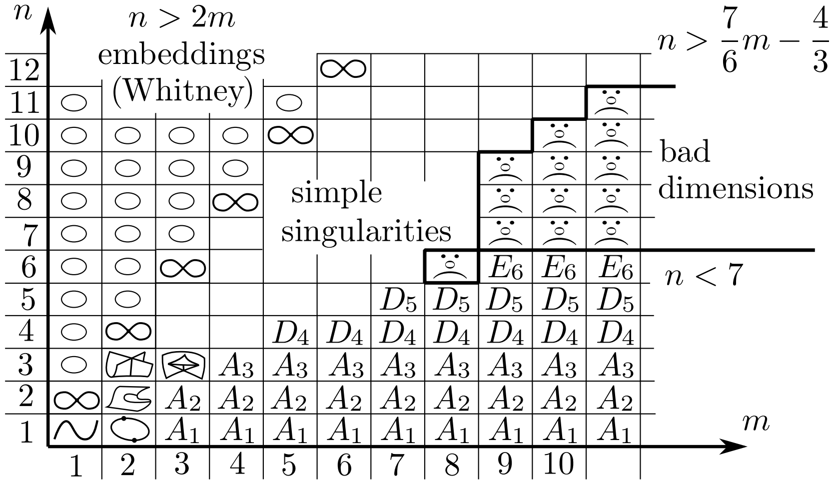


Figure 15.15: Good and bad dimensions for smooth generic maps $f : M^m \to N^n$.

(continuous invariants that distinguish the non-equivalent maps). If a pair of dimensions is bad, the stable maps do not form a dense set.

The effort of many mathematicians led to table 15.15. R. Thom was the first to prove that, in the cases $m = n$, the Whitney paradise of finite classifications of generic maps ends at $m = 9$, due to the presence of moduli for any $m \geqslant 9$. J. Mather made the definitive calculation of the whole boundary between the pairs of good and bad dimensions.

To explain the notations of the remaining (m, n) places, we have to mention the classification principle used to compose the table: "modality". The same principle is used to classify objects of different natures (varieties, functions, etc.), obtaining similar classifications and hierarchies in many cases.

For example, the classification of singularities in the domain of good dimensions is related to many other classifications: that of regular polyhedrons, of crystallographic reflection groups, of simple complex Lie algebras, of simple singularities of caustics and wave fronts, and so on.

15.7.1 Codimension vs Modality

Let us compare two classification principles. At first glance, the most natural classification principle is "up to codimension $\leqslant k$": it means to represent the entire space of objects (under consideration) as a finite union of submanifolds of codimensions not greater than k (called *classes*) and a remainder of codimension $> k$ so that within each class the objects' properties that interest to us do not change.

Example. For the germs of smooth functions of two variables with critical point 0 and critical value 0, the classification up to codimension $\leqslant 4$ is formed by the classes of functions that are locally transformable to functions of the following list (by a diffeomorphic change of independent variables):

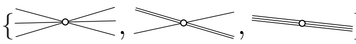
A_1	A_2	A_3	A_4	D_4	A_5	D_5
$\pm x^2 \pm y^2$	$x^3 \pm y^2$	$\pm x^4 \pm y^2$	$x^5 \pm y^2$	$x^2y \pm y^3$	$\pm x^6 \pm y^2$	$x^2y \pm y^4$

In general, classification up to codimension k differs from classification by codimensions of orbits (or of classes under some equivalence relation). The appearance of objects whose orbits have codimension k may turn out to be unavoidable in generic families with less parameters than k.

Example. In the classification of functions of two variables, the functions with a singularity of type $x^4 + y^4 + ax^2y^2$ form a set of codimension 7 (it is impossible to avoid the appearance of such singularities by a small perturbation of a generic 7-parameter family), but the codimension of the orbit of each function of this 1-dimensional set of functions is at least eight. Most functions of this family belong to different orbits and, hence, every neighbourhood of each function of this family contains a 1-parameter family of non-equivalent functions. The parameter a is thus a modulus (like the modulus of the holomorphic classification of elliptic curves, see p. 156).

k-tuples of lines. A classification problem in which the appearance of moduli is immediately seen is that of k-tuples of vectorial lines in the plane:

A. *Triples of Lines.* Consider the classification of the triples of straight lines that contain the origin of the plane $\mathbb{R}^2$, where the equivalence is defined by the linear transformations of the plane. Identifying the triples with the cubic forms vanishing on these lines, we easily prove that the classification is discrete: It consists of the three classes

$\{xy(x+y)\,,\ x^2y\,,\ x^3\}$ (that is, {  })

of respective dimensions 3, 2 and 1 in the 3-dimensional variety of triples.

B. *Quadruples of Lines.* Classifying the quadruples of straight lines containing the origin, up to linear transformations of the plane, we have the

Theorem 4. *The discrete classification of the quadruples of one-dimensional subspaces of the vector-space $\mathbb{R}^2$ is impossible. The set of equivalence classes is of positive dimension.*

Proof. The dimension of the variety of the quadruples is four. The dimension of the Lie group $\mathrm{GL}(2, \mathbb{R})$ is also four, but the dimension of the orbits is at most three because the scalar matrices fix every quadruple, and hence the orbits coincide with those of the 3-dimensional group $\mathrm{SL}(2, \mathbb{R})$. So, the space of orbits is at least 1-dimensional, and hence the discrete classification is impossible. Another proof is provided by the fact that the classes of quadruples are explicitly parametrised by the cross-ratio of the four lines. □

Modality. The *number of moduli* or *modality* of an object with respect to an equivalence relation (e.g. of a point $\omega \in V$ under the action of a Lie group G on a manifold V) is the least integer m for which there exists a neighbourhood (of that object) which can be covered by a finite number of families of orbits, whose number of parameters is at most m.

Example. Algebraically, the most natural classification of critical points of function germs is not by multiplicity or up to a given codimension, but by modality. In particular, functions with the same modality have other properties that are similar (their intersection forms, monodromy groups, etc.) and they appear together in other classification problems.

Simple Objects. Objects of modality zero are called *simple*: A neighbourhood of a simple object is covered by a finite number of orbits.

A germ of a function at a critical point is *simple* if it can be deformed in only a finite number of ways (up to smooth changes of the independent variables). The list of simple critical points of holomorphic functions of two variables consists of two series of singularities and three exceptional ones:

A_k, $k \geqslant 1$	D_k, $k \geqslant 4$	E_6	E_7	E_8
$x^{k+1} \pm y^2$	$x^2y \pm y^{k-1}$	$x^3 + y^4$	$x^3 + xy^3$	$x^3 + y^5$

(if the number of variables is larger, it is necessary to add a nondegenerate quadratic form in the missing variables). It coincides with the list of Weyl groups (the crystallographic reflection groups of Sect. 7.10) whose Dynkin diagram has no multiple edge. The connection between singularities and reflection groups is expressed in Th. 5 below and is discussed in Ch. 16.

Adjacencies. Since a neighbourhood of a simple object is covered by a finite number of classes (orbits), to which that object is "*adjacent*", these classes (orbits) form a discreet invariant of that simple object and have to be known.

The following adjacency diagram of the simple singularities of functions shows the classes to which a simple singularity is adjacent. Each arrow indicates the adjacency of a class of higher codimension (consisting of complicated singularities) to a larger class (consisting of less complicated singularities). All critical point types are adjacent to the Morse ones (of type A_1).

Adjacency diagram

$$\begin{tikzcd}
A_1 & A_2 \arrow[l] & A_3 \arrow[l] & A_4 \arrow[l] & A_5 \arrow[l] & A_6 \arrow[l] & A_7 \arrow[l] & A_8 \arrow[l] & \cdots \arrow[l] \\
 & & & D_4 \arrow[ul] & D_5 \arrow[l] \arrow[ul] & D_6 \arrow[l] \arrow[ul] & D_7 \arrow[l] \arrow[ul] & D_8 \arrow[l] \arrow[ul] & \cdots \arrow[l] \\
 & & & & & E_6 \arrow[ul] \arrow[uul] & E_7 \arrow[l] \arrow[ul] \arrow[uul] & E_8 \arrow[l] \arrow[ul] \arrow[uul] & \\
 & & & & & & (P_8) \arrow[ul] & (X_9) \arrow[ul] & (J_{10}) \arrow[ul]
\end{tikzcd}$$

A small variation of a function may transform its critical point into several critical points of those types to which the given critical point is adjacent (that is, of those types that are reachable through a path formed by arrows).

The codimension of a class is smaller by one than the index in the notation, which is equal to the multiplicity, that is, to the number of Morse critical points (of type A_1) colliding at the given critical point.

The three symbols in brackets do not represent classes, but three sets that contain all singularities more complicated than those on the list. They are all adjacent to E_6, E_7 or E_8 and form a set of codimension 6 in the space of all functions of $n > 2$ variables (codimension 7 for $n = 2$ and codimension ∞ for $n = 1$). Thus, a typical family of functions of n variables, depending on at most $k = 5$ parameters, contains no singularity other than the A_ℓ, D_ℓ, E_ℓ singularities (with $\ell \leqslant k + 1$). The simplest nonsimple singularity in the set P_8 is given by the nondegenerate cubical form in three variables

$$f_a(x, y, z) = x^3 + y^3 + z^3 + axyz.$$

Different values of the parameter a generically provide nonequivalent singularities because the corresponding elliptic curves $f_a = 0$ in $\mathbb{C}\mathrm{P}^2$ cannot be holomorphically transformed into one another (see p. 156).

15.7.2 Topological classification

In the bad dimensions there is no discrete classification of generic maps for the smooth equivalences, but the topological classification of maps, of caustics and of wave fronts remains discrete and finite.

The problem to classify these discrete topological equivalence classes of generic smooth maps is a great challenge, but no result or even no conjecture has been yet discovered. The difficulty arises already at the boundary between the good and the bad dimensions, like at the line $n = 7m/6 - 4/3$ of Fig. 15.15.

The singularities that are possible in good dimensions correspond (in a way not explained in this book, see for instance [115]) to discrete subgroups of the group SO(3) of rotations of the sphere $\mathbb{S}^2$. On the boundary of the domain of bad dimensions, one can expect singularities related (in the same way not explained in the book) to subgroups of motions of Euclidean plane, and therefore to the elliptic curves classification where a modulus appears. The following strange phenomenon occurs: There are some elliptic curves that are topologically distinguished and for which the corresponding values of the moduli should be considered as a separate strata of the topological classification.

It is difficult to compute these special values of the moduli of the elliptic curves. It seems that modern computers are still too weak for these topological problems.

Three evident examples of such special elliptic curves are related to the affine reflection groups of Euclidean plane generated by the triangles with angles $\left(\frac{\pi}{3}, \frac{\pi}{3}, \frac{\pi}{3}\right)$, $\left(\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{4}\right)$ and $\left(\frac{\pi}{2}, \frac{\pi}{3}, \frac{\pi}{6}\right)$. But there exist perhaps some dozens of other "topologically" exceptional elliptic curves, where some topological structures of the corresponding wave fronts and caustics bifurcate, whose calculations have frightened the experts.

15.8 Stability and Bifurcations

We have discussed the generic singularities of smooth maps and their classifications. But the natural approach suggested in the *Thèse* of Poincaré is different: Whenever a singularity is not generic, instead of the study of the properties and "normal forms" of these exceptional cases, Poincaré suggested to see first whether they are frequently needed. (Experience shows that, in most cases, the simplest and frequently encountered singularities have the greatest practical value; while in the more complicated cases, the expenditure of energy in surmounting technical difficulties is not always justified by the practical value of the obtained results.) So nongeneric objects may be neglected in an initial analysis. For example, the singular points of a generic vector field are nodes, foci and saddles, but not centres at all. Thus the investigation of centres may be put off as less important.

This reasoning of Poincaré lessens the value of so many traditional results of mathematical analysis in which special attention is given to the more difficult, but rarely encountered, degenerate cases.

15.8.1 Structural Stability

Of course, the study of (not so) degenerate cases is of highest interest if one does not consider an individual object, but a family of objects that depend on parameters (being the case in most concrete situations). But since the exact values of the parameters in a model are unknown, a conclusion drawn from the mathematical investigation of a model (in economy, physics, engineering, mathematics,. . .) is worth the confidence of the practical worker only to the extent that this conclusion is stable under small changes of the parameters. In general, considering the space of all the objects under study, the classification one uses is determined by the questions that one asks about a system, forbiding questions for which a small change of the model alters their answers. This led A.A. Andronov to the following concept.

Structural Stability. To have *structural stability* of properties of the system under study, the class formed by the objects equivalent to a given one (in relation to the properties of interest to us) must be open in the corresponding space of objects (functions, fields, bifurcations, etc.).

Depending on the classification being studied, the space Ω of all possible objects that we wish to investigate is subdivided into equivalence classes – Fig. 15.16. The necessity to investigate the open classes (consisting of stable objects, equivalent to their neighbours) is obvious.

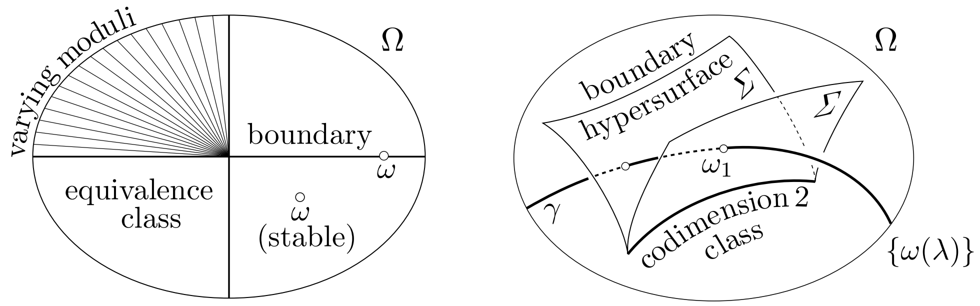


Figure 15.16: The bifurcation set Σ, formed by the non stable objects, decomposes the space Ω into open equivalence classes of stable objects.

15.8.2 Bifurcation Sets

There exist, however, a boundary hypersurface Σ that separates different open classes. It is called a *bifurcation set*. An object ω that belongs to this boundary is infrequent in the applications of our Ω-theory, as of a model, to real world phenomena because the parameters that define exactly the point ω in Ω are known practically only with some accuracy. Therefore, we can perturb ω slightly to make it stable but remaining, up to a given accuracy, at the same place ω.

The situation is different for a 1-parameter family of objects $\lambda \mapsto \omega(\lambda)$. Such a curve $\gamma = \{\omega(\lambda)\}$ can intersect the bifurcation set Σ transversely at an isolated point (which corresponds to a *bifurcation value* of the parameter, say, at $\omega_1 = \omega(\lambda_1)$) at which the class of the object $\omega(\lambda)$ changes: the object undergoes a "qualitative change" or "perestroika". Such transition from one type of singularity to another is a stable event: any neighbouring curve $\widetilde{\gamma}$ intersects the bifurcation set Σ at a neighbouring point $\widetilde{\omega}_1 = \widetilde{\omega}(\widetilde{\lambda}_1)$.

Thus, although the event of belonging to Σ is removable by a small deformation of the object ω, this event becomes unavoidable if instead of the individual object ω we wish to study the behaviour of the whole family (knowing, say, which change of the salary λ would lead to the revolution).

More degenerate objects whose equivalence classes form a higher codimension subvariety in Ω remain avoidable by small generic deformations, even in the study of one-parameter families, similar to the curve γ of Fig. 15.16 where a class of codimension 2 (of dimension $\dim \Omega - 2$, if $\dim \Omega < \infty$) is shown. Indeed, the curve γ could intersect this codimension 2 class, but a slightly generically deformed curve $\widetilde{\gamma}$ would certainly avoid it. Therefore, the codimension 2 events are a less necessary object of study than the codimension 1 events if one studies 1-parameter families.

These reasonings led Poincaré to a very important conclusion: Studying non generic objects, one has to calculate first the codimension of the corresponding event, say k, and then one has to study the bifurcations in generic k-parameter families, in which these degeneracies are unremovable, rather than the original unperturbed non generic object (say, $\omega(0, \dots, 0)$), whose study is insufficient in all real applications if it is not followed by the study of the behaviour of the perturbed objects in a generic family.

15.8.3 Bifurcation Diagrams

In sufficiently good cases, the bifurcation set Σ in the function space Ω (of all possible objects that we wish to investigate) has the structure of a local direct product of its section by a finite-dimensional subspace (transverse to the stratum of Σ to which the point being studied belongs) with an infinite-dimensional manifold of finite codimension (equal to the dimension of the cross-section and to the codimension of the stratum) along which "nothing essential changes".

In those cases, a generic family with finitely many parameters is a transverse section of the indicated stratum. The bifurcation set (in the function space) leaves a trace in the parameter space of the family, the preimage, called *bifurcation diagram* (formed by the bifurcation parameter values).

This bifurcation diagram has the same local structure as the trace of the *bifurcation set on the cross-section described above* (up to multiplication by a smooth manifold, if the number of parameters is larger than the codimension of the stratum).

The finger print of a singularity, which contains its most important informations, is its bifurcation diagram. In particular, the bifurcation diagram of a simple object geometrically exhibits the classes to which that object is adjacent, together with their disposition around that object. An example is the bifurcation diagram of zeros (compare Figures 5.8 and 5.9 of pp. 139-140 with Fig. 16.20 of p. 619 and Fig. 16.8 of p. 595).

Bifurcation diagrams of zeroes. The subset of functions having a critical value zero, in the space of all smooth functions, is called the *bifurcation set of zeroes.* It has a natural stratification in which different strata correspond to functions having different numbers of critical points of different types.

Since the space of functions may be considered as a (linear) surrogate of the space of hypersurfaces (each function represents its zero level hypersurface), the bifurcation set of zeroes can be seen as the subset of singular hypersurfaces in the space of all the hypersurfaces.

The trace of the bifurcation set of zeroes on the parameter space of a family of functions is called the *bifurcation diagram of zeroes of the family.*

Example. In Fig. 5.8, p. 139, we have described the bifurcation diagram of zeroes of the two-parameter family of functions of x (for the singularity A_2)

$$f(x) = x^3 + ax + b\,,$$

which is the semicubic parabola along which the discriminant $4a^3 + 27b^2$ vanish. Similarly (for A_3) the bifurcation diagram of zeroes of the family

$$g(x) = x^4 + ax^2 + bx + c\,,$$

described in Fig. 5.9, p. 140, is the swallowtail surface discussed in p. 555.

A generic family intersects the strata of simple singularities (A, D, E) transversely and, for such a family, the singularity of the bifurcation diagram of zeroes is the same at every point of a given simple stratum (up to a biholomorphic diffeomorphism and up to multiplication by a smooth space if the number of parameters is greater than the codimension of the stratum). In particular, the bifurcation diagram is independent of the number n of variables of the functions in our family (which number might be even infinite) and of the choice of a generic family.

Theorem 5. *The bifurcation diagram of the zeros of a generic family of functions is biholomorphically equivalent to the discriminant of the corresponding finite reflection group* (multiplied by a smooth space if the number of parameters is greater than the codimension of the stratum) *at a neighbourhood of each point of each simple singularity stratum A, D, E* (see Sect. 7.10).

There is no evident relation between the classifications of the Weyl groups of simple Lie algebras and of the simple singularities of functions. The proof is based on a comparison of the two independent classifications and on a comparison of the discriminants with the bifurcation diagrams.

Thus, the bifurcation diagrams are universal hypersurfaces with singularities (independent from the family if it is generic) in finite-dimensional spaces, which reveal that *few simple general laws govern many very different phenomena*. In particular, from the large number of different 'classification' problems in singularity and bifurcation theory there is a comparatively small list of standard forms (the semi-cubic cusp, the swallowtail, the Whitney umbrella, etc.) serving in different theories that have no apparent connection.

Remark. While the bifurcation diagram of a family separates its real parameter space into parts, formed by the objects which are "qualitatively the same" (e.g. the real swallowtail of p. 140), a complex hypersurface does not separate a complex parameter space into parts. In the complex case, the separating role of the real bifurcation diagram is transformed into that of the branching set for the monodromy, which is the natural action on our objects of the fundamental group of the complement to the bifurcation diagram (e.g. the fundamental group of the complement to the complex $(n-1)$-dimensional swallowtail hypersurface in $\mathbb{C}^n$ is the braid group $\mathrm{Br}(n+1)$ - see p. 140).

15.8.4 Choosing the Classifying Group

While an "uncorrect" classification principle (e.g. by codimensions) leads to a chaotic and difficult to understand hierarchy of increasingly complicated bifurcation diagrams, the "correct" classification (e.g. by modality) is governed by simple general rules.

Many other examples, in the investigation of singularities, show that an unfortunate choice of the equivalence relation leads to a difficult-to-handle chaos, while a nearby, but fortunate, choice may lead to a simple and beautiful, clearly final, classification.

For example, the classification of the perestroikas of wave fronts by the so called (r, s)-equivalence led to a complicated and difficult-to-handle classification [136], while in the "correct" setting [13], after doing only a small change of the equivalence group, the results immediately become natural and easy to handle.

Unfortunately, there is no general rule to select successful formulations of classification problems, and they must be found gropingly among many variants that seem equally worthy. However, the following two principles have proved to be very fruitful:

(*i*) *One may consider a classification more successful, the larger is the number of different kinds of problems in which it appears.*

For example, the A, D, E - classification (of Sect. 7.10 and of p. 560), appears in contexts of such different natures as the theories of simple singularities, of crystallographic reflection groups, of simple Lie algebras, of simple representations of quivers and of regular polyhedra in 3-dimensional Euclidean space (where E_6 corresponds to the tetrahedron, E_7 to the octahedron, and E_8 to the icosahedron).

(*ii*) *To compare different classifications one should begin with the investigation of bifurcation diagrams of objects of low codimension.*

The coincidence of bifurcation diagrams (serving as distinctive fingerprints of singularities) in various classification problems allows one to establish natural isomorphisms between hierarchies of objects of different theories. Thus the coincidence of bifurcation diagrams has led to the discovery of the connection between singularities of functions on a manifold with boundary, and the Weyl groups B, C, F (whose Dynkin diagram has one double edge, p. 263). For example, the F_4 caustic illustrated in Fig. 16.10 is realised in the theory of boundary singularities (see p. 599).

15.9 Versal Deformations

The study of the perturbed objects (discussed above) was called by Poincaré *bifurcation theory*. This continuation of singularity theory is extremely important in many applied domains. Poincaré himself applied his bifurcation theory to the study of the periodic solutions of dynamical systems, especially in celestial mechanics. Poincaré proved his unexpected theorems, in the case of analytic functions, using an important general fact of holomorphic calculus, which he called "Lemma 4". The further generalisation of Lemma 4 is called "the versal deformation theorem". It was formulated by R. Thom and proved by B. Malgrange, but we don't like to forget the Poincaré "Lemma 4", which, in modern terminology, is "the analytic versal deformation theorem for complete intersections".

From the previous sections we can conclude that the best way to understand the behaviour and properties of a degenerate object is to perturb it inside a generic family that is large enough to give all essentially distinct bifurcations of the given object and such that the singularity type in question does not vanish away under a small deformation of the family (for a codimension k event we need to dispose of at least k parameters). Such a family is called "versal family" or "versal deformation".

To explain the versal deformations, we shall describe below their application to a simple problem of linear algebra: In the study of normal forms of the linear operators that send a vector-space to itself, $A : \mathbb{C}^n \to \mathbb{C}^n$, we can see that the well known theory of Jordan normal forms cannot resist the Poincaré critics because a generic matrix has no multiple eigenvalues.

Thus, since the events of multiple eigenvalues are degenerate, their study ought to start with the calculation of the codimension, say k, of the corresponding varieties in the space $\mathbb{C}^{n^2}$ of matrices of order n. Next one ought to investigate generic k-parameter perturbations of the given matrix. The versal deformation theorem provides normal forms for these deformations.

Definition. A k-parameter *deformation* of an element $v \in V$ is a smooth map $F : B \subset \mathbb{R}^k \to V$, such that $F(0) = v$, where the *parameter space* (or *base*) B is a neighbourhood of $0 \in \mathbb{R}^k$ (of $0 \in \mathbb{C}^k$ in the complex theory).

One can think on the submanifold $F(B) \ni v$ as the deformation of v.

Definition. Consider a smooth action $A : G \times V \to V$ of a Lie group G on a manifold V. Two deformations F and $\widetilde{F}$ with base B are *equivalent with respect to* G or G*-equivalent*, if there exists a smooth map $g : B \to G$ such that $g(0) = e$ and

$$A\big(g(b), F(b)\big) = \widetilde{F}(b).$$

It means that the two images $F(b)$ and $\widetilde{F}(b)$ of each $b \in B$ belong to the same orbit under the action of G and, hence, that they can be moved to one another by an element $g(b)$ of the group. (The condition $g(0) = e$ means that the map $g : B \to G$ is a k-parameter deformation of e in G.)

Example. Let V be the space of the $n \times n$-matrices of linear operators in $\mathbb{R}^n$ (or $\mathbb{C}^n$) and $G = \mathrm{GL}(n, \mathbb{R})$ (or $\mathrm{GL}(n, \mathbb{C})$) the group of the non degenerate matrices, which act on V as changes of coordinates:

$$A(g, v) = g \cdot v \cdot g^{-1},$$

where the dots mean the multiplication of matrices. In this case, the equivalence of the deformations $\{F(b)\}$ and $\{\widetilde{F}(b)\}$ of a matrix $v = F(0) = \widetilde{F}(0)$

means the conjugation of the matrices $\widetilde{F}(b)$ to the matrices $F(b)$ by a coordinate change $g(b)$ that depends smoothly on the parameter $b \in B$:

$$\widetilde{F}(b) = g(b) \cdot F(b) \cdot g^{-1}(b) \ , \quad \text{where } g(0) = e \ .$$

Definition. A deformation $\widetilde{F}$ with base space $\widetilde{B}$ is said to be *induced from the deformation* F with base space B, if there exists an "inducing" smooth map $\varphi : (\widetilde{B}, 0) \to (B, 0)$ such that $\widetilde{F} = F \circ \varphi$.

In particular, given a deformation $F : (B, 0) \to (V, v)$, any smooth map $\varphi : (\widetilde{B}, 0) \to (B, 0)$ induces a deformation $\widetilde{F} : (\widetilde{B}, 0) \to (V, v)$ from F just by the composition of maps: $\widetilde{F} = F \circ \varphi$.

Example. The deformation $x^3 + \varepsilon^2 x$ with parameter ε of the function x^3 is induced from the deformation $x^3 + \lambda x$ with parameter λ by the map $\varphi(\varepsilon) = \varepsilon^2$.

Versal Deformations. A deformation F of a point v is said to be *versal* if any other deformation of v is G-equivalent to a deformation induced from F by some smooth map φ.

The word "versal" is the intersection of the words "universal" and "transversal". The versal deformation is universal in the sense that it "contains" all the arbitrary deformations up to the natural equivalence relation: To study all the possible events in any deformation of v, it suffices to know what happens for some particular values of the parameter b in the versal deformation.

The transversality will appear soon in the following proof of the existence of the versal deformation, which is not evident at all.

Versal deformation existence. We start by considering the submanifolds $M_v \subset V$ and $H_v \subset G$ associated to a point $v \in V$ (see Fig. 15.17):

M_v – *The orbit of the point* v is the submanifold $M_v = A(G \times \{v\}) \subset V$.

Write N for the dimension of V and $m = \dim V - \dim M_v$ for the codimension of the orbit $M_v^{N-m} \subset V^N$.

H_v – *The stationary subgroup of* v is the following submanifold of G:

$$H_v = \{h \in G : A(h, v) = v\} \ .$$

Denote by L the dimension of the group G. The dimension of the submanifold H_v is equal to $L - \dim M_v$ by the implicit function theorem because the image of the tangent space $\mathfrak{g}$ of G under the linearised map A_{*e} is the tangent space to the orbit M_v at v.

The dimension of the orbit M_v is $N-m$ and, hence, any complementary transverse subspace to the tangent space

$$T_eH = (\operatorname{Ker}(A_* : T_eG \to T_vV))$$

in T_eG has also dimension $N-m$.

Deformation construction – Let D^m be a smooth m-dimensional submanifold of V^N, containing the point v, whose tangent space at v is transverse to the tangent space of the orbit M_v:

$$T_vD^m + T_vM_v^{N-m} = T_vV^N \ , \quad (T_vD^m) \cap (T_vM_v^{N-m}) = 0 \ .$$

We shall prove that the submanifold D^m, transverse to the orbit M_v^{N-m} of v, (more precisely, its inclusion $D^m \to V$) is a versal deformation of v.

Let E^{N-m} be a smooth $(N-m)$-dimensional submanifold of the group G, transverse to the stationary subgroup H_v of the point v:

$$T_eE^{N-m} + T_eH_v^{L-(N-m)} = T_eG \ , \quad (T_eE^{N-m}) \cap (T_eH_v^{L-(N-m)}) = 0 \ .$$

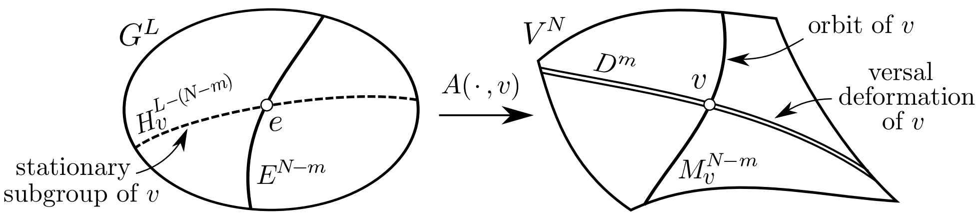


Figure 15.17: Construction of a versal deformation D of the point v.

Consider the smooth map $Q : \left(E^{N-m} \times D^m\right) \to V^N$, defined in the neighbourhood of the point (e, v) by the group action of the elements of the submanifold E^{N-m} on the points of the submanifold D^m of V^N:

$$Q(\varepsilon, \delta) = A(\varepsilon, \delta) \ , \quad \varepsilon \in E^{N-m} \ , \delta \in D^m \ .$$

Proposition. *The map Q is a local diffeomorphism.*

Proof. First, the restriction of Q to the submanifold $\{e\} \times D^m$ (identified with D^m) is the identity map, since $A(e, \delta) = \delta$ for any $\delta \in D^m$. Hence, the derivative of Q sends the tangent space of D^m to itself as the identity map.

Next, the tangent space of E^{N-m} is sent onto the tangent space of the orbit M_v^{N-m} because it does not intersect the nonzero part of the tangent space $T_e H = \mathrm{Ker}\,(A_*)$.

Therefore, the image of the derivative Q_* at (e, v) covers the whole tangent space $T_v V$ and, hence, Q is a local diffeomorphism. □

It follows that, in some neighbourhood of the point v in V, each element w has a unique representation in the form

$$w = A(\varepsilon, \delta) \ ,$$

where $\varepsilon = \varepsilon(w)$ and $\delta = \delta(w)$ depend smoothly on w.

This is the versality statement for the deformation D^m of v. Indeed, the diffeomorphism property of Q implies that for any deformation $\widetilde{F}$ each deformed point $w = \widetilde{F}(\widetilde{b})$ of $v = \widetilde{F}(0)$ has the representation

$$\widetilde{F}(\widetilde{b}) = A\big(\varepsilon(\widetilde{F}(\widetilde{b})), \widetilde{\delta}(\widetilde{F}(\widetilde{b}))\big) \ .$$

This means that the map that sends $\widetilde{b}$ to $\delta(\widetilde{F}(\widetilde{b}))$ induces from the deformation D^m a deformation equivalent to $\widetilde{F}$: The conjugating deformation of e is $g(b) = \varepsilon(\widetilde{F}(\widetilde{b})))$.

Remark. Our versal deformation of the element $v \in V$ is m-parametric, where m is the codimension of the orbit of v.

It is clear that any versal deformation cannot have less parameters. For this reason such a deformation is called *miniversal*.

It is in general not *uni*versal, since the universality definition (omitted above) requires the unicity of the inducing map φ, which would be a very exceptional situation (in general, the miniversal deformation intersects some neighbouring orbits at many points). The (not defined) universality property fails if there exists a diffeomorphism of the parameter space of the deformation sending any point to another point of the same orbit. As a rule, such diffeomorphisms exist unless every orbit is represented on the parameter space by at most one point.

Remark. In the main applications of the versal deformation theorem to singularity theory, the space V is infinite-dimensional and one deals with infinite-dimensional groups G of diffeomorphisms. In these cases, the implicit function theorem is not available because the orbit M and the "manifold" D are also infinite-dimensional.

However, in spite of these difficulties, versal deformations exist in many cases and the proofs can avoid the references to untrue theorems of abstract functional analysis.

Example. For the R-equivalence the miniversal deformation of the function x^3 is the two-parameter deformation $x^3 + \lambda x + \mu$ (for the LR-equivalence it is $x^3 + \lambda x$).

Similarly, the codimension of the orbit of the function x^n under the action of the group of right equivalences is $m = n - 1$. The gradient ideal contains all functions divisible by x^{n-1} and its R-miniversal deformation is

$$x^n + \lambda_1 x^{n-2} + \lambda_2 x^{n-3} + \cdots + \lambda_{n-1}$$

(to be compared with the Whitney normal forms).

In the classification of families of matrices discussed above, the implicit function theorem works well, and we get the following explicit formulae for the versal deformations. Here the group $\mathrm{GL}(n, \mathbb{C})$ acts on any complex $n \times n$ matrix v as $g \cdot v \cdot g^{-1}$.

Theorem 6. *The codimension m of an orbit in the space of complex $n \times n$ matrices is given by*

$$m = \sum_{\lambda} (n_1 + 3n_2 + 5n_3 + \dots) \ ,$$

where $n_1 \geqslant n_2 \geqslant n_3 \geqslant \dots$ are the sizes of the Jordan blocks of the matrix that have the same eigenvalue λ and the coefficient of n_k is $2k - 1$.

Theorem 7. *The miniversal deformation $v + B$ of a Jordan matrix v with a unique eigenvalue λ can be chosen in the form shown in* Fig. 15.18 (for the case of three Jordan blocks of sizes $n_1 \geqslant n_2 \geqslant n_3$).

In Fig. 15.18, we represent an $n \times n$ matrix B, where $n = n_1 + n_2 + n_3$. Each element of this matrix is zero, except the elements of the Jordan blocks and the $n_1 + 3n_2 + 5n_3$ elements staying at the stressed places, which are the m independent parameters of the versal deformation.

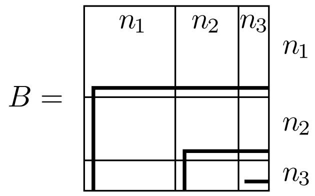


Figure 15.18: The versal deformation of the Jordan matrix with eigenvalue λ, having 3 Jordan blocks of sizes $n_1 \geqslant n_2 \geqslant n_3$.

Example. The matrix of order 5 with two Jordan blocks of eigenvalue α of sizes 3 and 2 has the versal deformation with nine parameters b_j:

$$\begin{pmatrix} \alpha & 1 & 0 & 0 & 0 \\ 0 & \alpha & 1 & 0 & 0 \\ b_1 & b_2 & \alpha + b_3 & b_4 & b_5 \\ b_6 & 0 & 0 & \alpha & 1 \\ b_7 & 0 & 0 & b_8 & \alpha + b_9 \end{pmatrix}$$

For the case of several eigenvalues, the versal deformation should deform independently the diagonal blocks that correspond to each eigenvalue, following the prescription of Theorem 7.

The proofs of Theorems 6 and 7 are published in [9]. Some applications of the resulting normal forms of the families of matrices to such problems, like the generic bifurcations of the boundary of stability and the behaviour of the matrix increments (singularities of $\max_k\{\lambda_k(v+B)\}$ as a function of the parameters B) are published in [15], pp. 238-260.

It follows, for instance, that in the generic k-parameter families of matrices only a finite number of different singularities of the stability boundary $(\max(\mathrm{Re}\ \lambda_k) \leqslant 0)$ occurs, up to local diffeomorphisms of the parameter space. For example, for $k = 3$ the number of these singularities is 4: the dihedral angle, the trihedral angle, one half of the Whitney umbrella and the surface $X^2Y^2 = Z^2$, $X \geqslant 0$, $Y \geqslant 0$.

Structural and Deformational Stabilities. In the contemporary theory of singularities, instead of the function space of all objects, one immediately considers generic families with finitely many parameters, and, instead of describing a neighbourhood in the function space of the object being studied, one studies families with finitely many parameters that contain this object. Thus, the structural stability of an object (a whole neighbourhood belongs to the same equivalence class) is replaced by deformational stability (each deformation of the object with finitely many parameters is trivial).

Similarly, structural stability of a deformation of an object (each nearby family is locally equivalent to the family which gives the deformation) is replaced by versality (each deformation of the same object with finitely many parameters is equivalent to a deformation induced from the given one).

For most classifications in Singularity Theory the deformational stability of an object implies its structural stability, and versality implies the structural stability of a deformation. However, these facts are not evident a priori, and their proof in concrete classifications is generally not simple.

15.10 Topological mathematics

The "topological approach" to many domains of mathematics is an informal idea that cannot be reduced to the replacement of the diffeomorphisms by the homeomorphisms in the definitions of the equivalences, of the invariants and in the classifications.

For example, all members f_ε of the family of functions

$$f_\varepsilon(x) = x^3 + \varepsilon^2 x \ ,$$

are topologically equivalent (say, for the R-topological classification).

However, the value $\varepsilon = 0$ is clearly different from the others, and it should be distinguished in any reasonable classification.

In this particular example, the reducing homeomorphisms are no longer possible in the complex domain. However, in other cases, the complex objects may be homeomorphic for all values of the parameters involved in their definition, but being "exceptional" for some special values of the parameters that would not be distinguished by the topological equivalence approach.

For example, the complex semicubic parabola $x^2 = y^3$ in the affine plane $\mathbb{C}^2$ is homeomorphic to $\mathbb{C}$, but should be distinguished from $\mathbb{C}$ in any reasonable discrete classification.

Hence, we prefer the following informal definition of topology, instead of the usual "study of the properties invariant under homeomorphisms":

Topology is the study of discrete features of structures and of discrete invariants in any domain of continuous mathematics, independently of the topological invariance by homeomorphisms.

We would prefer the name " X-adjective Topology" for the domain of any science X that studies those discrete features and invariants.

Thus, the Möbius theorem stating that "*any curve obtained from a smooth perturbation of a straight line in* $\mathbb{R}\mathrm{P}^2$ *has at least three inflection points*" belongs to projective topology.

We would like to include the studies of the eigenvalue multiplicities (like the Wigner-Neumann theorem on the repulsion of the eigenvalues of the symmetric matrices), the eigen-frequencies monotonicity theorem and the eigenvalues intermittence theorem for constrained oscillatory systems into linear-algebraic topology (or topology of linear algebra).

The necessity of at least four cusps on the *caustic* of a generic convex closed curve C in Euclidean plane (i.e. the curve obtained as the envelope of the normal straight lines to C) belongs to symplectic topology, as well as the "symplectic ribs", obstructing the symplectic camel percolation from one part of the 4-space to the other through a small hole in the three-dimensional wall separating them.

The Plücker formulas, relating the numbers of singular points of different kinds on a projective plane complex algebraic curve of degree n and the invariants of its projectively dual curve, belong to the topology of algebraic geometry.

The maximum principle for the harmonic functions and the index calculation for boundary problems in the theory of partial differential equations belong to topologic mathematical physics.

The Gauss-Bonnet formula and the Gauss linking number formula both belong to the topology of differential geometry.

Pontryagin generalised the Euler characteristic, equal to the sum of the indices of the singular points of a vector-field on a smooth manifold, to the higher degrees tensor-fields, obtaining new remarkable numerical conditions of the coexistence of several different degenerations of a smooth tensor-field on a given smooth manifold.

These new characteristic numbers and characteristic classes are not always invariant under the homeomorphisms, but they provide the most interesting results of topology (distinguishing, for instance, the 28 homeomorphic non-diffeomorphic spheres of Milnor, described on page 110).

There is no doubt that the theories of these characteristic classes and characteristic numbers form the very content of the topology of analysis, both when they are topologically invariant (like the Euler characteristic) and when they are not.

The Newton integrability theorem (saying that "all Abelian integrals along conic section curves are elementary functions") and the Newton non-integrability theorem (claiming that "the area of the real plane segment between a curve and a chord is a non elementary function of the chord position, whatever were the closed curve") both these theorems, as well as their proofs, belong to the topology of complex analysis.

Indeed, the elementariness depends on the fact that the genus of the Riemannian surface of a circle is equal to zero, while the non-integrability is a corollary of the infiniteness of the monodromy group of the area function of the segment.

There is a lot of open problems in all these directions of topological sciences. The invariance by homeomorphism is weakly related to the heart of those topics. Of course, the discrete invariants, which cannot change continuously with the parameters (like can, say, the eigenvalues), are in many cases provided by the "genuinely topological" invariants of the homeomorphisms.

However, in many cases the actions of the homeomorphisms are undefined in the theory X that we wish to topologise. Neither symplectomorphisms in symplectic topology, nor contactomorphisms in contact topology, nor cusps in algebraic and projective topology have "topological" (homeomorphisms invariant) versions. In spite of it, Lagrangian intersection theory, Legendrian link theory and Floer homology are topological theories.

Some mathematicians attempted to define X-topology as the study of those properties of objects of (smooth) theory X, which are continuous functions of the smooth objects in C^0-topology (for instance, whose values remains constant under the smooth deformations of small magnitudes, even if the deformations of the derivatives are not small).

Such a definition might replace formally the continuation of the invariance by diffeomorphisms to the invariance by homeomorphisms, which is undefined in the smooth theory X, leading to many theorems and especially to lots of Ph. D. dissertations.

However, J. Sylvester explained many years ago that the axiomatic petrification of any mathematical idea is a (tempting) danger of the modern science, since the formalised version is always weaker, and some of the most promising applications are eliminated by the axiomatisation of the informal idea. In his words, mathematical ideas should be preserved, like the flow of the river: You never arrive twice to the same water, but the current is still the same [121] .

Concerning the topological classification of generic smooth maps in the bad dimensions, we shall explain the difficult problem of the distinction of the special elliptic curves for the cases of the first bad dimensions.

One of the objects associated to the singularity type is a pair formed by the corresponding wave front (discriminant hypersurface) and the caustic (bifurcation hypersurface).

The real (or the complex) discriminant hypersurface of dimension $N-1$, is a part of $\mathbb{R}^N$ (or of $\mathbb{C}^N$). It is naturally projected to the space $\mathbb{R}^{N-1}$ (respectively, to $\mathbb{C}^{N-1}$). The caustic hypersurface in $\mathbb{R}^{N-1}$ (in $\mathbb{C}^{N-1}$) is the projection of the $(N-2)$-dimensional "cuspidal edge" of the discriminant hypersurface to the space $\mathbb{R}^{N-1}$ (respectively, to $\mathbb{C}^{N-1}$).

An example of this situation is depicted in Fig. 15.14: The projection of the discriminant surface $f(\Sigma) \subset \mathbb{R}^3$ to the plane $\mathbb{R}^2$, along the vertical c-direction, generates a caustic curve that has a semicubic cusp in the plane $\mathbb{R}^2$ with coordinates a and b. It is the projection of the cuspidal edge $f(\Sigma')$ of the swallowtail surface $f(\Sigma)$.

In the bad dimensional case, where moduli are present, all these objects (discriminant, projection, caustic) depend on parameters, say, on the elliptic curve type, which is a smooth invariant, under diffeomorphisms, of the singularity type.

The problem is to understand whether the topological structure of the projection of the discriminant and of its cuspidal edge to $\mathbb{R}^{N-1}$ (or $\mathbb{C}^{N-1}$) changes or persists under the small variations of the parameter values.

The general theory implies that these topological changes have a discrete structure: There is a finite set of special values of the moduli of the elliptic curves at which the topological structure varies, but it is locally constant at the other places.

The difficulty is, however, that up to now no one has been able neither to calculate the number of these special elliptic curves nor to give an explicit description of them.

We leave aside the problem of the possible discrete invariants that generalise the topological invariance definition: The *topological* structure of the projection of the discriminant and of its cuspidal edge to the base $\mathbb{R}^{N-1}$ (or to $\mathbb{C}^{N-1}$) of the fibration of $\mathbb{R}^N$ (or $\mathbb{C}^N$) into "vertical lines" is already a new structure that might be modified by the homeomorphisms, transforming one of the topologically equivalent maps $f : M^m \to N^n$ to the other, $\widetilde{f}$ (while the diffeomorphisms that realise the equivalence of f to $\widetilde{f}$ would preserve the topological structures of the discriminant, of the caustic and of the projection).

15.11 Complex singularities

In this chapter, we have mainly considered singularities of maps between real manifolds. However, it is well known that, going to the complex case, mathematical problems usually are simplified. For example, every algebraic equation of degree d has exactly d complex roots, while determining the number of real roots is a difficult problem. The reason of this phenomenon, and of many others, is the "Italian Principle" discussed and used in Ch. 5.

This chapter was mainly to devoted to singularities of maps between real manifolds topics for which, on the whole, it was not important which field (real or complex) we were considering. Hoever, there are questions for which it is important to work in the complex domain; for example, the decomposition of singularities, the connection between singularities and Lie algebras and the asymptotic behaviour of different integrals depending on parameters become clearer in the complex domain.

And there are topics for which the study of real singularities and complex singularities lead to theories, techniques and results (monodromy).

A comparison of the c1assifications of complex and real singularities reveals that all zero- and unimodal real singularities are real forms of the corresponding complex singularities. This fact is not a priori obvious and emerges only on comparing the independently carried out real and complex classifications.

The point is that it is not known whether or not modality is preserved under complexification. There are examples of a representation of a real Lie group for which the modality of a point increases under complexification (E.B. Vinberg). For critical points the modality cannot decrease under complexification (V.V. Muravlev), but it is not known ifit can actually increase. Recently V.V. Serganova and V.A. Vasil'ev have constructed an example of a real singularity whose proper modality (i.e., the dimension of the $\mu = \text{const}$ stratum in the base of the miniversal deformation) is lower than the proper modality of its complexifica- tion, see V.A. Vasil'ev, V.V. Serganova.

Monodromy. Morse theory studies the restructurings, perestroikas, or metamorphoses that the level set $f^{-1}(x)$ of a real function $f : M \to \mathbb{R}$, defined on a manifold M, undergoes as X passes through the critical values of f. The Picard-Lefschetz theory is the complex analogue of Morse theory. In the complex case the set of critical values does not divide the range $\mathbb{C}$ of a complex-valued function into connected components, and no restructurings occur: all level manifolds close to a critical one are topologically identical.

For this reason, in the complex case, rather than passing through a critical value, one has to go around it in the plane $\mathbb{C}$ where the function takes its values.

If we fix a small circle that goes around the critical value, then to each point of the circle there corresponds a nonsingular level manifold of the function. The set of all such levels is a fibration over the circle. Going around the circle defines a map of the homology of the fibre [over the initial point] of this fibration into itself. This map is called the monodromy corresponding to the critical value of the singularity. It is precisely the monodromy that represents the complex analogue of the restructurings in Morse theory.

****

AVG-II :

In the topological investigation of isolated critical points of complex-analytic functions the problem arises of describing the topology of its level sets. The topology of the level sets or infra-level sets of smooth real-valued functions on manifolds may be investigated with the help of Morse theory (see [255]). The idea there is to study the change of structure of infra-level sets and level sets of functions upon passing critical values. In the complex case passing t1!rough a critical value does not give rise to an interesting structure, since all the non- singular level sets near one critical point are not only homeomorphic but even diffeomorphic. The complex analogue of Morse theory, describing the topology of level sets of complex analytic functions, is the theory of Picard-Lefschetz (which historically precedes Morse theory). In Picard-Lefschetz theory the fundamental principle is not passing through a critical point but going round it in the complex plane. Let us fix a circle, going round the critical value. Each point of the circle is a value of the function. The level sets, corresponding to these values, give a fibration over the circle. Going round the circle defines a mapping of the level set above the initial point of the circle into itself. This mapping is called the (classical) monodromy of the critical point. The simplest interesting example in which one can observe all this clearly and carry through the calculations to the end is the function of two variables given by

Dehn twist, vanishing cycles, Picard Lefschetz formula

Milnor's book

José Seade's book: a collection of essays on selected topics on the topology and geometry of real and complex isolated singularities.

As we know, the bifurcation diagrams of zeros of the simple singularities are diffeomorphic to the varieties of non-regular orbits of the corresponding reflection groups. This fact permits the use of techniques developed in the theory of reflection groups in investigations of the bifurcation diagrams of simple singularities.

Huygens, without the help of analysis, was able to solve the majority of problems solved by Newton; but such solutions required the genius of Huygens. Nowadays, the same problems may be solved by any student with the help of analysis. In the same way, the techniques of singularity theory allows one to obtain results automatically and to investigate with less effort, and more rigour, more complicated singularities for which "elementary" methods would lead to vast calculations (requiring inventiveness and substantial efforts).

These ideas were systematically used in thermodynamics from the time of J.C. Maxwell and, especially, J.W. Gibbs. The perestroika (Fig. 22) of the isotherms of van der Waals' equations of state is a typical example of an application of the geometry of the pleat. In the time of Maxwell it was well known that this geometry is independent of the exact form of the equations of state.

Physicists systematically applied singularity theory before it was born. In the hands of L.D. Landau, the art of throwing away inessential terms of Taylor series, retaining the higher-order, but "physically important" terms, yielded many excellent results, now included within catastrophe theory.

Chapter 16

Symplectic and contact geometry and topology

Symplectic and contact geometry are the product of the evolution of variational calculus, dynamical systems (especially Hamilton systems of classical mechanics), geometrical optics, wave propagation theory, quasiclassical asymptotics in quantum mechanics, microlocal analysis of PDEs and Lie theory of diffeomorphism groups and Poisson algebras.

One of the most successful applications of Singularity Theory has been the study of singularities and bifurcations of caustics and wave fronts, which are determined by the singularities that appear generically in symplectic and contact geometry (the Legrangian and Legendrian singularities discussed below). For example, last years, the ideas, techniques and results on singularities of wave fronts and caustics revealed to be a powerful tool to discover new and natural theorems on the differential geometry of curves and surfaces.

16.1 What is symplectic geometry about?

In this section, we review some relevant results of symplectic geometry and topology in a survey style. The proper definitions are given in the next section, where we explain all the necessary to enter into the subject.

From Hamilton dynamics to symplectic geometry

The geometrisation of mathematics and physics, originating in the pioneering works of Poincaré, has led to the description of the time evolution of the state of a dynamical system in terms of a flow of the so-called phase fluid in the phase space, whose points represent different states of the system (say, the positions and the velocities of its particles).

The phase flow consists of transformations g^t of the phase space, depending on time, sending any initial state to the new state of the system after time t. The transformations forming the phase flow of a Hamilton dynamical system are not arbitrary diffeomorphisms. For example, they cannot have attractors by Liouville theorem, which says that the phase flow of a Hamiltonian system preserves the volume of any domain of the phase space.

One can thus imagine that the phase fluid filling the phase space is incompressible. This incompressibility property has several consequences (for example, the Poincaré return theorem and the numerous applications of ergodic theory to dynamical systems) that lead to consider the "volume-preserving geometry", which is the study of geometric properties of different objects on a manifold with a fixed volume element (here, *geometric* means invariant under the volume-preserving diffeomorphisms). Statistical mechanics is based on this incompressibility of phase flows.

Symplectic Geometry arose from the understanding that the transformations forming the phase flows of dynamical systems in classical mechanics and in variational calculus (and hence in optimal control) belong to a narrower class of diffeomorphisms of the phase space than the incompressible ones. Such diffeomorphisms, called *symplectomorphisms*, preserve a closed non degenerate differential 2-form ω called the *symplectic structure* of the phase space. (The integral of ω along a 2-surface in the phase space, called the *Poincaré integral invariant*, is preserved by the phase flows of Hamilton dynamical systems.)

The symplectomorphisms form a group and have peculiar geometric and topological properties which are responsible for many astonishing facts in mechanics, optics, and other parts of mathematical physics. While the first examples of such phenomena were explicitly described by Poincaré (and might be traced to the works of Hamilton, Jacobi, Lie, Cayley, and others), the systematic study of the geometry on symplectic manifolds is mostly due to the works of the second half of the 20th century. Applications to statistical physics (where the physicists freely permute the pieces of the phase space, provided that they have equal volumes) are still to be found.

Symplectic geometry is the geometry of symplectic manifolds and maps. Symplectic topology is younger (see [3], [21]). Let us review some results.

Symplectic Camel Problem. The conjecture that symplectomorphisms may behave differently than volume-preserving diffeomorphisms was first formulated as follows. Consider the "eye of the needle" (a hole in a vertical plane in three-space). In volume-preserving geometry any camel, however large it is, can snake from one half-space to the other through the hole – Fig. 16.1.

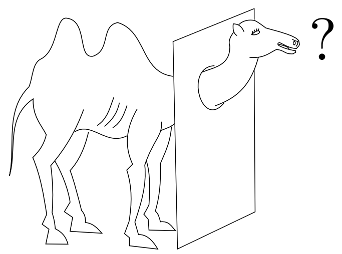


Figure 16.1: The symplectic camel problem.

The symplectic camel problem asks *whether it is possible for a symplectic camel in one half of the symplectic space to be continuously transformed into a camel in the other half, passing through a small hole in the hyperplane separating the half-spaces.* The transformation should be defined by time-dependent symplectomorphisms of the complement of the hyperplane with a hole. These diffeomorphisms should connect the identity map with the map sending the camel from one half space to the other.

Any symplectic space is even-dimensional. Symplectic geometry in dimension two coincides with the volume (area)-preserving geometry. The two-dimensional camel can percolate through any hole from the left half-plane to the right one. In dimension four, however, a symplectic camel has *symplectic ribs* that do not permit him to pass from one half-space to the other if the hole is not large enough (see [64], [99], [135]).

These "ribs" are defined in terms of the periods of periodic solutions of some Hamiltonian differential equations associated to the camel. The impossibility of traversing can be considered as an extension from linear to nonlinear oscillations of the Rayleigh-Fisher-Courant theorem on the behaviour of the eigenvalues under imposed constraints.

The Symplectic Packing Problem. Any bounded domain of the plane, whose area is smaller than the area of the unit disc, can be sent into this disc by an area-preserving diffeomorphism. A similar result holds in the volume-preserving geometry in any dimension.

The symplectic packing problem requires one *to send via symplectomorphism a bounded domain of the symplectic space into a given bounded domain of larger volume.*

It is impossible, for instance, to embed by a symplectomorphism the unit ball of the standard symplectic four-space into the product of two two-dimensional symplectic discs, one of which has radius $r < 1$ (see [76]).

A collection of disjoint images of k equal balls (fig. 16.2) under a system of symplectomorphisms of symplectic four-space can fill no more than $1-1/N$ of the volume of the symplectic four-ball, and can fill almost $1-1/N$ of the volume, where N is equal to $2, 4, \infty, 5, 25, 64, 289, \infty$ for the respective values of k: $2, 3, \ldots, 9$ ([100], [49]). Moreover, for every positive integer

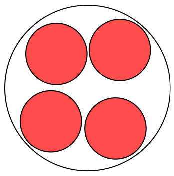


Figure 16.2: The symplectic packing problem.

p the symplectic $2n$-ball can be filled with gaps of arbitrary small volume by the disjoint images of p^n equal symplectic balls ($N = \infty$, see [100]).*

The universal obstacles in the symplectic embedding and packing problems are called *Gromov's width and symplectic capacities* (see [63] and [84]).

Last geometric theorem of Poincaré

The first theorem of symplectic topology was discovered by Poincaré in his studies of periodic orbits in celestial mechanics. It is called *the last geometric theorem of Poincaré*, because he had announced it just before his death, being unable to prove it. The proof was given later by Birkhoff.

Poincaré had formulated his theorem for the area-preserving maps of an annulus to itself: *such a map has at least two fixed points, provided that it moves the two boundary circles in opposite directions*. From the modern point of view this is a consequence of some general facts of symplectic topology.

Symplectic Fixed-Point Theorem. Consider a two-dimensional torus with its area element as a symplectic manifold. A symplectomorphism is called *exact*, if it can be connected to the identity map by a continuous path in the group of symplectomorphisms "preserving the centre of masses" of the torus. This "preservation" condition can be written in terms of a coordinate system on the plane covering the torus as follows – Fig. 16.3.

*However, for any $k > 9$ there exists a symplectic filling of the 4-ball by k symplectic 4-balls, leaving empty only an arbitrarily small part.

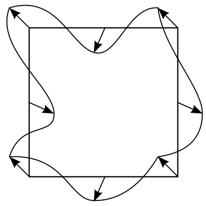


Figure 16.3: An exact symplectomorphism of a torus.

If the coordinates are (x, y) mod 1, the area element is $dx\,dy$ and the map sends the point (x, y) to $(x + f(x, y), y + g(x, y))$, then the centre-of-mass preservation condition requires the vanishing of the average shift, that is, of the integral of the vector (f, g) along the unit square. The theorem (see [60]) says that *any exact symplectomorphism of a torus* has at least four fixed points, and at least three of them are geometrically different.

Remark. The original statement of Poincaré follows because one can construct a torus by gluing together two identical annuli. The theorem says that *an exact diffeomorphism of the torus has at least as many fixed points as a function on the torus has critical points.* In this form, the theorem holds for all surfaces and for many higher-dimensional manifolds.

Example. On a surface of genus g the number of fixed points counted with multiplicities is at least $2g+2$, and at least 3 among them are geometrically different. For the torus T^{2n} these numbers are 2^{2n} and $2n+1$, respectively. For the complex projective space $\mathbb{C}\mathrm{P}^n$ both numbers are $n+1$. The proofs for some classes of symplectic manifolds are in [60], [76], [70], [82], [72], etc.

The theorem is still neither proved nor disproved for arbitrary compact symplectic manifolds (the exact symplectomorphisms form the commutator subgroup of the connected component of the identity in the symplectomorphism group). However, in the beginning of the 21st century, several research teams claimed to have proved the theorem in its general form. In fact, what they have proved is that *the number of fixed points is less than or equal to the sum of the Betti numbers of the manifold.*

Although the Betti inequalities are always satisfied:

$$\#\,(\text{critical points}) \;\geqslant\; \sum_i b_i\,,$$

most part of the known manifolds satisfy the equality $\#$ (critical points) $= \sum_i b_i$. However, it is possible to construct manifolds for which the strict inequality holds, # (critical points) $> \sum_i b_i$. For such manifolds the conjecture has not been proved.

The Morse numbers and the critical points occur in this problem by no coincidence. The whole theory can be viewed as an extension of Morse theory to generalised multivalued functions, called "Legendrian manifolds".

Infinitesimal exact symplectomorphisms are defined by Hamiltonian vector fields. The critical points of the Hamiltonian are fixed points of these symplectomorphisms. The symplectic fixed-point theorem extends the Morse-theoretic minoration of the number of fixed points of infinitesimal exact symplectomorphisms to the exact symplectomorphisms themselves. Many other facts of classical calculus may also be considered as infinitesimal versions of theorems in symplectic geometry.

Topology of caustics and wave fronts

Four-vertex theorem. Consider a convex curve in Euclidean plane (Fig. 16.4). A vertex is a local extreme point of the curvature radius. The classical four-vertex theorem says that *the number of vertices is at least four*, [104].

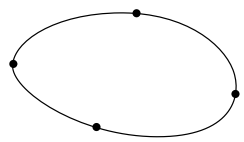


Figure 16.4: The vertices of a plane curve.

It seems a theorem of Riemannian geometry, but it does not hold for the geodesic curvature of curves on the plane even with a Riemannian metric close (but distinct) to the Euclidean one. Construct an example!

The real meaning of this theorem can only be understood in terms of symplectic and contact topology, where it appears together with quite a few other global theorems on singularities of caustics and wave fronts.

An infinitesimal counterpart of these general theorems is the Sturm-type result of p.345 which implies that any 2π-periodic function of the form $f'''+f'$ has at least 4 zeroes on every period.

Focal points. A *focal point* of a point on a surface equipped with a Riemannian metric is an intersection point of a geodesic ray issued from the point with an infinitely close geodesic ray also issuing from this point.

The set of focal points of a given point of a generic surface form a curve, called the *caustic* of the initial point. The caustic of the North pole of a sphere consists of its South pole and North pole. This caustic is completely degenerate, due to the high symmetry of the sphere. Perturbing the metric of the sphere (say, transforming it into an ellipsoid) one transforms the caustics of its points into small but complicated curves. Jacobi proved that the caustic of any point on a convex surface has cusps – Fig. 16.5.

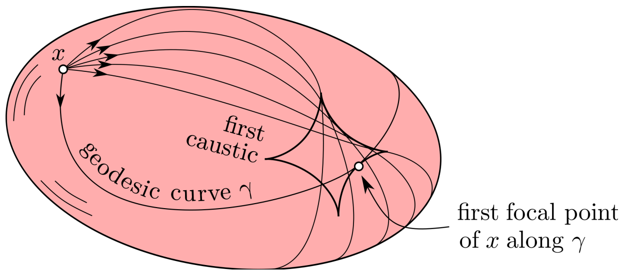


Figure 16.5: The first caustic of a point on a surface.

The caustic consists of several branches. The first one consists of the first focal points along each ray; the second of the next ones, and so on.

If the perturbation is small, the first branch of the caustic of the North pole is close to the South pole, the second to the North pole and so on. The smallness of the perturbations one needs depends of the order of the branch.

The four cusp theorem says that *the caustic* (say) *of a generic point on a generic convex surface has at least four cusps*.

This is classically proved for the first branch of the caustic. It holds also for any branch, provided that the perturbation is sufficiently small (see [28]). Standard conjectures in symplectic topology imply that it should be true with no smallness restrictions, but this is not proved.

16.2 Symplectic manifolds

Symplectic manifolds are the natural generalisations of the phase spaces of classical mechanics, that is, of the cotangent bundle spaces.

A *symplectic manifold* is a manifold equipped with a closed non degenerate differential 2-form ω, called *symplectic structure* or *symplectic form*.

Symplectic manifolds are even dimensional because any skewsymmetric form in an odd-dimensional space is degenerate.

Example. The plane with its oriented area element is a symplectic manifold.

Example. The space $\mathbb{R}^{2n}$ has a natural symplectic structure. Its value on a pair of vectors is the sum of oriented areas of the parallelograms constructed with their projections on the n coordinate 2-planes (p_i, q_i) : $\omega = \sum dp_i \wedge dq_i$.

Example $\mathcal{N}$. The most important examples of symplectic manifolds are the cotangent bundle spaces of manifolds (like the phase space of a classical mechanical system, whose base manifold is called *configuration space*).

A *cotangent vector* on a manifold is a linear function on the space of tangent vectors at a point of the manifold. The set of all cotangent vectors (called *momentum vectors* in physics) at all points of a base manifold B^n form the *cotangent bundle* T^*B (of dimension $2n$).

The *action* 1- *form* (or *Liouville form*) on the cotangent bundle is the tautological one: it is the form α whose value on any vector ξ tangent to the cotangent bundle at a point $p \in T^*_x B$ is equal to the value of the cotangent vector p on the projection of this tangent vector to the base manifold: $\alpha(\xi) = p(\pi_* \xi)$ (Fig. 16.6). The *natural symplectic structure* of the cotangent bundle is the 2-form $\omega = d\alpha$.

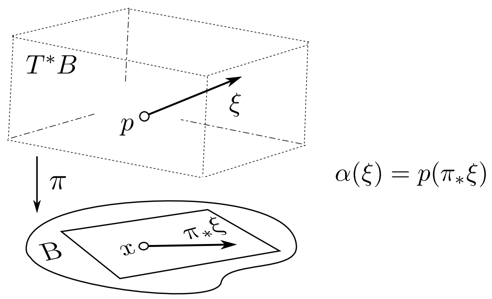


Figure 16.6: The action form on the cotangent bundle.

For example, the action 1-form on the cotangent bundle of the vector space with linear coordinates $(q_1, \ldots, q_n)$ is $\alpha = \sum p_i \, dq_i$ (where the p_i are the natural coordinates of a cotangent vector) providing $\omega = \sum dp_i \wedge dq_i$.

More Examples. 1. The Kähler manifolds* of complex geometry; 2. The orbits of the coadjoint representation of any Lie group in the dual space of its Lie algebra; 3. The complex projective space. 4. The manifold of

*A *Kähler manifold* is a complex manifold which also carries a Riemannian metric and a symplectic structure on the underlying real manifold in such a way that the three structures (complex, Riemannian, and symplectic) are all mutually compatible.

all complex matrices with fixed and simple eigenvalues. 5. The Cartesian product of two symplectic manifolds.

PROBLEM. Prove that the manifold of all oriented lines in $\mathbb{R}^n$ is symplectic.

Hint. This manifold may be identified with the total space of the (co)tangent bundle of the sphere $\mathbb{S}^{n-1}$. Associate to each oriented line its unit velocity vector $v \in \mathbb{S}^{n-1}$. The point of intersection of the line with the hyperplane of $\mathbb{R}^n$ tangent to $\mathbb{S}^{n-1}$ at v defines the required element of $T^*\mathbb{S}^{n-1}$.

A *symplectomorphism* is a diffeomorphism preserving the symplectic structure. All differences between symplectic manifolds are global (similar to Euclidean spaces but unlike Riemannian manifolds):

Darboux Theorem. *All symplectic manifolds of any given dimension are locally symplectomorphic to one another.*

Corollary (**Darboux Coordinates**)**.** *In a sufficiently small neighbourhood of every point of any symplectic manifold, the symplectic structure can be expressed in appropriate local coordinates in the form* $\omega = \sum dp_i \wedge dq_i$.

Moreover, any point of a connected symplectic manifold can be carried to any other by a flow of symplectomorphisms.

In symplectic geometry, in contrast to Riemannian geometry, there exist no local exterior invariants of submanifolds:

Darboux-Givental Theorem ([38])**.** *The restriction of the symplectic structure to a submanifold of a symplectic manifold defines this submanifold locally in a neighbourhood of any of its points up to a local symplectomorphism.*

Example. Any two smooth hypersurfaces (that is, defined locally by one nondegenerate equation each) in a symplectic manifold are locally symplectomorphic in the neighbourhoods of any two points.

The algebraic (or better analytic) geometry of symplectic manifolds and of their subvarieties should be based on the interaction of two algebra structures in the space of functions: The ordinary commutative multiplication and the Poisson bracket Lie algebra structure. Many facts which are known in the regular case of transversal submanifolds, should extend to the general situation. A motivation for this extension is the following remark due to Melrose (who has based on it his works on diffraction singularities near gliding rays):

The geometry of submanifolds of a Riemannian manifold can be considered as the symplectic geometry of pairs of submanifolds in a symplectic manifold.

Example. The differential geometry of a surface $F(q) = 0$ in Euclidean space can be read from the symplectic geometry of a pair of hypersurfaces in phase space.

Namely, the equation $F(q) = 0$ also defines a hypersurface in the cotangent bundle of Euclidean space (the *phase space*) and the Riemannian metric can be viewed as the hypersurface $p^2 = 1$ of momenta of length one in the phase space.

16.2.1 Characteristics, Hamilton fields, Poisson bracket

Characteristic Direction. All points of a smooth hypersurface in a symplectic manifold are equivalent and at each point of the hypersurface there exists a unique preferred tangent direction intrinsically defined by the symplectic structure. This *characteristic direction* is the skew orthocomplementary direction of the tangent hyperplane of the hypersurface.

Hamilton Field. The *Hamiltonian vector field* of a function H on a symplectic manifold is the field X_H for which the value of the symplectic structure on a pair formed by a vector of the field and any second tangent vector v of the symplectic manifold at the same point is equal to the derivative of the function along the second vector (pp. 239-240), $\omega(X_H, v) = dH(v)$.

The function is then called the Hamiltonian of the field. It is defined by the field up to a locally constant summand. The Hamiltonian vector field is the only vector field intrinsically associated to a function on a symplectic manifold (up to multiplication by a function locally constant on the level sets of the Hamiltonian function).

Example. A hypersurface in a symplectic manifold is foliated (locally fibred) into its *characteristics* – curves tangent to the vectors of the Hamiltonian fields defined by all functions constant along the hypersurface (which are independent of the choice of the equation defining the hypersurface).

Hence the structure of the decomposition of the hypersurface into characteristics (which may be very complicated as is well known from celestial mechanics and Hamiltonian chaos theory) is a symplectic-topological invariant of the hypersurface. It is also a source of symplectic invariants of domains bounded by hypersurfaces in symplectic space.

The Hamiltonian vector fields form a Lie algebra which is a sub-algebra of the algebra of all the vector fields with the usual *Poisson bracket* operation.

The Hamiltonian functions form a Lie algebra whose operation is also called Poisson bracket. The *Poisson bracket of two functions* is the derivative of one of them along the Hamiltonian vector field of the other. The algebra of functions is a central extension of the algebra of fields (whose centre consists of the locally constant functions).

The flow of every Hamiltonian vector field preserves the symplectic structure. In return, if a flow preserves the symplectic structure, then it is locally a Hamiltonian field, but globally its Hamiltonian function may be multivalued. The Lie algebra of Hamiltonian vector fields is a sub-algebra of the algebra of *symplectic* vector fields (those preserving the symplectic structure).

Example. The translations of the torus preserve the area element. In consequence, the constant (translation-invariant) vector fields on the torus are symplectic. However they are not Hamiltonian vector fields.

The flows of Hamiltonian vector fields on the torus are precisely those consisting of area-preserving and centre-of-mass preserving diffeomorphisms. They form the commutator sub-algebra of the Lie algebra of symplectic vector fields. The quotient space is the one-dimensional cohomology space of the symplectic manifold.

16.3 Lagrangian Submanifolds

The geometry of a symplectic manifold is refreshingly different from the usual Euclidean or volume-preserving geometry. Some of its submanifolds are locally different from others of the same dimension.

Definition. A submanifold of a symplectic manifold (M, ω) is called *isotropic* if the restriction to it of the symplectic form ω vanishes. The isotropic submanifolds of maximal dimension (which is n in a symplectic manifold of dimension $2n$) are called a *Lagrangian submanifolds.*

Example. Any smooth curve on the plane is a Lagrangian submanifold.

Example. For any function $q \mapsto S(q)$, $q \in \mathbb{R}^n$, the submanifold $\mathcal{L}$ of the standard symplectic space $\mathbb{R}^{2n}$ defined by $p = \partial S/\partial q$ is Lagrangian. Indeed, on this submanifold $pdq = dS$ and hence the restriction of $dp \wedge dq$ to $\mathcal{L}$ vanishes. The function S is called *generating function* of $\mathcal{L}$.

In general, any smooth function f defines a Lagrangian section of the cotangent bundle of its domain of definition $x \mapsto p = df_x$.

Example. The fibres of any cotangent bundle $T^*B \to B$ are Lagrangian submanifolds. Indeed, the standard 1-form $\alpha = pdq$ vanishes along the fibres, implying that its differential $\omega = d\alpha$ also vanishes.

Problem $\mathcal{N}$. Let N be a submanifold in Euclidean space $\mathbb{R}^n$. Prove that the n-dimensional manifold L_N formed by the covectors $\langle v, \cdot \rangle$ at the end-points of the normal vectors v to N is a Lagrangian submanifold of $T^*\mathbb{R}^n$.

Solution. Under the Euclidean identification of tangent and cotangent vectors, the set of all vectors normal to N form a submanifold $\widetilde{L}$. Since the restriction of the standard 1-form $\alpha = pdq$ in $T^*\mathbb{R}^n$ to this submanifold vanishes, $\widetilde{L}$ is Lagrangian. The map $T^*\mathbb{R}^n \to T^*\mathbb{R}^n$, $(p, q) \mapsto (p, p+q)$ preserves the symplectic structure and sends $\widetilde{L}$ onto L_N. Therefore L_N is Lagrangian.

Example. Let N be a submanifold in Euclidean space $\mathbb{R}^n$. The set of oriented lines of $\mathbb{R}^n$ normal to N is a Lagrangian submanifold of the symplectic space of oriented lines in $\mathbb{R}^n$ (see problem of p. 589). Prove it!

According to the general principle of A. Weinstein "*the most important objects in symplectic geometry**, *after the symplectic manifolds themselves, are the Lagrangian manifolds*". For example:

A symplectomorphism $f : M \to N$ between two symplectic manifolds is ("materialised" as) a Lagrangian submanifold of the Cartesian product $M \times N$.

Indeed, equipping the product with the symplectic structure provided by the difference of (the pull-backs of) the two given structures, $\omega = \omega_N^* - \omega_M^*$ (and observing that a diffeomorphism is completely determined by its graph), we have that *the graph of a diffeomorphism $f : M \to N$ is a Lagrangian submanifold of the product $M \times N$ if and only if the map f is symplectic.*

Remark. Replacing here the graph by *any* Lagrangian submanifold of the product we obtain an interesting generalisation of the notion of symplectomorphism: *symplectic correspondence*. It is also useful to consider *Lagrangian varieties* with singularities.

Locally all Lagrangian submanifolds are symplectomorphic to each other.

Weinstein Theorem ([137]). *Some neighbourhood of any Lagrangian submanifold in any symplectic manifold is symplectomorphic to some neighbourhood of this Lagrangian submanifold in any other symplectic manifold, for instance in its own cotangent bundle space.*

In general, there is no such global symplectomorphism. This is clear from the case of plane curves: the area bounded by a closed curve is a symplectic invariant of the curve.

---

*In October 2009, at Université de Bourgogne, Weinstein told to Uribe-Vargas that now he simply says " in Geometry, every important object is a Lagrangian manifold".

16.4 Lagrangian Fibrations and Caustics

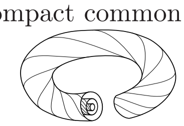
Lagrangian singularities are the singularities of Lagrangian maps, of which the simplest example and model is the projection of a Lagrangian submanifold in phase space to the configuration space, $L \subset T^*B \xrightarrow{\pi} B$ (in which all fibres of π are Lagrangian submanifolds, as we have seen).

Lagrangian Fibration. A fibration of a symplectic manifold is said to be *Lagrangian* if all its fibres are Lagrangian submanifolds.

Example. The cotangent bundle is a Lagrangian fibration. In particular, the so-called *standard Lagrangian fibration* $\mathbb{R}^{2n} \to \mathbb{R}^n$, $(p, q) \mapsto q$.

Problem. Prove that the map from the symplectic manifold of oriented lines in Euclidean space $\mathbb{R}^n$ to the unit sphere $\mathbb{S}^{n-1}$ associating to any oriented line its unit orienting vector, is a Lagrangian fibration.
Hint. It coincides with the cotangent bundle of the sphere $\mathbb{S}^{n-1}$.

Theorem 1. *All Lagrangian fibrations of the same dimension are locally* (*in the bundle space*) *symplectomorphic* (hence they are locally symplectomorphic to the cotangent bundle of Euclidean space).

Remark. The fibres of a Lagrangian fibration (or foliation) have natural *local affine structures*, intrinsically defined by the symplectic geometry of the fibration. This is crucial for the theory of integrable Hamiltonian systems and is very close to the Liouville theorem on integrable systems, which claims that the smooth compact common level manifolds of commuting Hamiltonian functions are nested tori ([2]) :  .

Lagrangian Map. Given an immersed Lagrangian submanifold $i : L \to E$ in the space of a Lagrangian fibration $\pi : E \to B$, its projection to the base of the fibration along the fibres, $\pi \circ i : L \to B$, is called a *Lagrangian map*.

Example $\mathcal{N}$. Consider the set of all vectors normal to a submanifold N in Euclidean space $\mathbb{R}^n$, say an ellipsoid in $\mathbb{R}^3$. To each vector v based at a point q of N associate its end-point $q + v$. This Lagrangian map of the Lagrangian submanifold L_N of $T^*\mathbb{R}^n$ (see Problem $\mathcal{N}$ p. 592) into Euclidean space $\mathbb{R}^n$ is called the *Normal Map* of the submanifold N, $L_N \to T^*\mathbb{R}^n \to \mathbb{R}^n$.

Example. Given a transversely oriented hypersurface M in Euclidean space $\mathbb{R}^n$, transport its orienting unit normal vectors to the origin. This defines a map to the unit sphere $\mathbb{S}^{n-1}$, $\Gamma : M \to \mathbb{S}^{n-1}$, called the *Gauß map* of M. It is the composition of the map that associates to each point of the hypersurface M the oriented normal line to M at that point (its image is a Lagrangian submanifold L_G in the space of oriented lines of $\mathbb{R}^n$, which is isomorphic to $T^*\mathbb{S}^{n-1}$) with the projection $T^*\mathbb{S}^{n-1} \to \mathbb{S}^{n-1}$, $M \to L_G \subset T^*\mathbb{S}^{n-1} \to \mathbb{S}^{n-1}$.

Since a Lagrangian map $\pi \circ i : L \to B$ is a map between two manifolds of the same dimension, it is a local diffeomorphism at most points of L, but at some points the rank of the differential drops. Under the projection of these singular points to the base space an “apparent contour” is formed :

Caustic The *caustic* of a Lagrangian map is the set of its critical values.

Optical caustic. In optics, if a submanifold N of Euclidean space $\mathbb{R}^n$ is (considered as) a source of light, its *caustic* or *focal set* is defined as the envelope of its normal lines (which are called *normal light rays*). Light intensity is much more concentrated on the caustic than in any other part of space.

If N is a hypersurface, its focal set consist of its centres of curvature.

The focal set of a generic curve $\gamma : \mathbb{R} \to \mathbb{R}^n$ has a rather simple description :
The hyperplane normal to γ at a point is the union of all lines normal to γ at that point. The envelope of all hyperplanes normal to γ is thus the focal set (for $n \geqslant 3$ the curve γ itself is also a degenerate component of the focal set, but we will not consider it).
The normal hyperplanes at two neighbouring points of γ intersect along an affine subspace of codimension 2 which approaches a limiting position as the points move into coincidence. The subspace that assumes this position is called the 2-*codimensional polar subspace* of the curve at the considered point. The union of these polar subspaces along the curve is, by construction, the envelope of the hyperplanes normal to γ, i.e. it is the focal set.

Example $\mathcal{N}$. The caustic of the normal map of a submanifold in Euclidean space $\mathbb{R}^n$ coincides with its *focal set*. For example, the caustic of the normal map of an ellipse in Euclidean plane is a curve, affine equivalent to an astroid, whose four cusps correspond to the four vertices of the ellipse (Fig. 16.7).

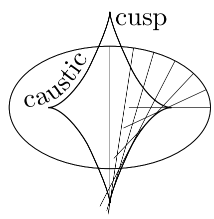


Figure 16.7: The caustic of an ellipse.

The 4-vertex theorem guarantees at least four cusps on the caustic of a generic convex curve. General results of symplectic geometry show that it is not a peculiarity of *Euclidean geometry*: If we start from a curve in the plane with another Riemannian metric, the Lagrangian variety is well defined and its caustic still generically has at least four cusps.

Example. The caustic of the Gauß map of a generic surface in Euclidean space $\mathbb{R}^3$ is the image of its parabolic curve under the Gauß map, because the Gauß map derivative is degenerate at the parabolic points (see p.318).

A *Lagrangian equivalence* of two Lagrangian maps is a symplectomorphism of the total space transforming the first Lagrangian fibration to the second, and the first Lagrangian immersion to the second. The equivalence classes of germs of Lagrangian maps are called *Lagrangian singularities*.

Caustics of equivalent Lagrangian maps are diffeomorphic.

Caustics can have complicated singularities; but (as in usual singularity theory) we can get rid of the overly complicated singularities by a small deformation of the Lagrangian submanifold under which it remains Lagrangian.

After this there remain only the simplest unremovable singularities, which we can study once and for all, and for which we can write normal forms (see the next section). For problems in general position we can expect that only these simple unremovable singularities of caustics will appear:
The only singularities of 1-dimensional caustics encountered in generic practical situations are semi-cubic cusps (Fig. 16.7) and transverse self intersections. All singularities of generic caustics in a 3-space are shown in Fig. 16.8.

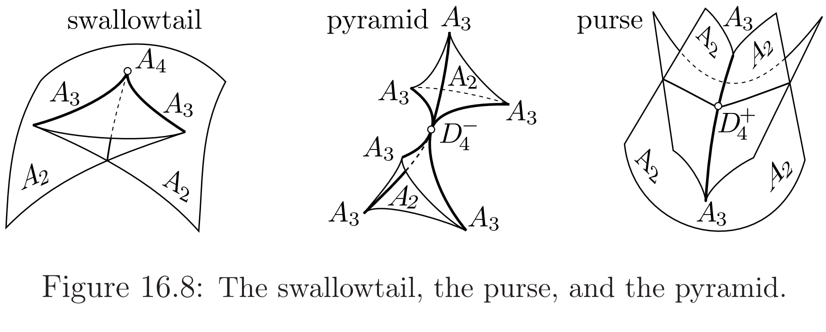


Figure 16.8: The swallowtail, the purse, and the pyramid.

Let us consider the model used to classify Lagrangian map singularities.

16.5 Generating Families of Lagrangian Maps

Consider the standard Lagrangian fibration $\mathbb{R}^{2n} \to \mathbb{R}^n$, $(p,q) \mapsto q$ with the symplectic form $dp \wedge dq$. Let $F(x,q)$ be the germ, at the point (x_0, q_0), of a family of smooth functions of k variables $x = (x_1, \ldots, x_k)$ which depends smoothly on the parameters $q \in \mathbb{R}^n$, such that at our point (x_0, q_0)

$$(i)\ \frac{\partial F}{\partial x} = 0 \quad \text{and} \quad (ii) \text{ The map } \ (x,q) \mapsto \frac{\partial F}{\partial x} \ \text{ has rank } k.$$

Then *the germ of the set* $L_F = \left\{(p,q) : \exists x : \frac{\partial F}{\partial x} = 0, p = \frac{\partial F}{\partial q}\right\}$, *at the point* $(\frac{\partial F}{\partial q}(x_0, q_0), q_0)$, *is the germ of a smoothly immersed Lagrangian submanifold of* $\mathbb{R}^{2n}$. The family germ F is called a *generating family* of the Lagrangian submanifold L_F and of the Lagrangian map $\pi_F : L_F \ni (q,p) \mapsto q$.

It turns out that *the germ of each Lagrangian map is equivalent to the germ of the Lagrangian map* π_F *for a suitable family* F.

The *caustic* of a (generating) family of functions consists of the parameter values for which the corresponding function has a non-Morse critical point.

Problem. Prove that the focal set of any generic curve $\gamma : \mathbb{R} \to \mathbb{R}^n$ is the caustic of the family of functions $F : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ given by $F(q;\vartheta) = \frac{1}{2} \parallel q - \gamma(\vartheta) \parallel^2$. That is, F is a generating family of the normal map of the curve γ.

Solution. The caustic of the family $F(q;\vartheta) = \frac{1}{2} \parallel q - \gamma(\vartheta) \parallel^2$ is the set of parameter values $q \in \mathbb{R}^n$ for which the function F_q has a degenerate critical point $\vartheta \in \mathbb{R}$, that is,

$$F_q'(\vartheta) = 0 \ \text{ and } \ F_q''(\vartheta) = 0. \tag{1}$$

For each value of ϑ, the points $q \in \mathbb{R}^n$ satisfying the first equation of (1) form the hyperplane normal to γ at $\gamma(\vartheta)$: $F_q'(\vartheta) = -\langle q - \gamma(\vartheta), \gamma'(\vartheta)\rangle = 0$.

The points $q \in \mathbb{R}^n$ satisfying both equations (1) for a fixed ϑ are the stationary points of the normal hyperplane at $\gamma(\vartheta)$ for infinitesimal variation of ϑ. Then they form the 2-codimensional polar subspace of the curve at $\gamma(\vartheta)$ (see p. 594), which is defined by the equations $\langle q - \gamma(\vartheta), \gamma'(\vartheta)\rangle = 0$ and

$$F_q''(\vartheta) = -\langle q - \gamma(\vartheta), \gamma''(\vartheta)\rangle + \langle \gamma'(\vartheta), \gamma'(\vartheta)\rangle = 0.$$

For $n = 2$ the additional condition $F_q'''(\vartheta) = 0$ provides the cusps of the caustic.

The classification of generic singularities of caustics in spaces of dimension $n < 6$ is known and is discrete ([10]). They are classified by the simple Lie algebras, A_μ, D_μ, E_μ (where $\mu \leqslant n+1$) and are equivalent to the *Lagrangian singularities* of the following list of generating families:

$A_\mu: \ F(x,q) = \pm x^{\mu+1} + q_1 x^{\mu-1} + \cdots + q_{\mu-1}x, \quad \mu \geqslant 1;$
$D_\mu: \ F(x,q) = x_1^2 x_2 \pm {x_2}^{\mu-1} + q_1 {x_2}^{\mu-2} + \cdots + q_{\mu-2}x_2 + q_{\mu-1}x_1, \quad \mu \geqslant 4;$
$E_6: \ F(x,q) = x_1^3 \pm x_2^4 + q_1 x_1 x_2^2 + q_2 x_1 x_2 + q_3 x_2^2 + q_4 x_1 + q_5 x_2;$
$E_7: \ F(x,q) = x_1^3 + x_1 x_2^3 + q_1 x_1^2 x_2 + q_2 x_1^2 + q_3 x_1 x_2 + q_4 x_2^2 + q_5 x_1 + q_6 x_2;$
$E_8: \ F(x,q) = x_1^3 + x_2^5 + q_1 x_1 x_2^3 + q_2 x_1 x_2^2 + q_3 x_2^3 + q_4 x_1 x_2 + q_5 x_2^2 + q_6 x_1 + q_7 x_2.$

From dimension 6 on, the classification of Lagrangian singularities up to symplectomorphisms (or even of singularities of caustics up to the diffeomorphisms) is no longer discrete. The *topological classification* is discrete in any dimension, but it is unknown even in dimension six.

16.5.1 Unimodular Singularities and Mirror Symmetry

The non discrete classification of the Lagrangian singularities of higher dimensions remains algebraically appealing, at least at the beginning, when they depend on few parameters. The classification of *unimodular singularities* (i.e., whose classes depend on one parameter) consists of several series of one-parameter families of singularities and of 14 sporadic families of unimodular singularities that correspond to the 14 exceptional triangles on the Lobachevsky plane. These 14 triangles have angles π/p, π/q, π/r, where (p,q,r) is one of the 14 triples

$$(239)(247)(336)(256)(345)(444)(238)(246)(335)(255)(344)(237)(245)(334)\,.$$

A strange duality between the 14 unimodular singularities was discovered long ago (see [11]). It is an involution on the set of 14 triangles which permutes the so-called *Gabrielov numbers* and *Dolgachev numbers* of the same singularity. V. Arnold called this duality *strange* because it is hard (impossible?) to guess which triangle is dual to which. Try to find the rule knowing the answer: the dual pairs are

$$(239) \leftrightarrow (334), (247) \leftrightarrow (335), (238) \leftrightarrow (245), (256) \leftrightarrow (344)\,.$$

Each of the remaining six triangles is dual to itself. The sum of all the six Gabrielov and Dolgachev numbers equals 24 for any of the 14 singularities. It was later discovered that this strange duality is a manifestation of the so-called *mirror symmetry* of three-folds studied by physicists. It is perhaps the first manifestation of this symmetry, which permutes the Hodge numbers of different-dimensional cohomologies and which itself is strange and rather poorly understood.

More general cases of mirror symmetries for hypersurfaces in the so-called toric varieties of arbitrary dimension (not only of dimension 3 of physicists, see [47]) are manifestations of projective duality of convex geometry applied to the Newton polyhedra.

The theory of mirror symmetry for the toric varieties themselves is closely related to the symplectic fixed point and Lagrangian intersections theories. The multiplication in Floer-type cohomology, studied in the framework of symplectic topology by Givental (see [81]), is now called *quantum cohomology* and even just "cohomology" by physicists.

16.5.2 Reflection Groups and Caustics

The classification of Lagrangian singularities remains discrete in higher dimensions in the neighbourhoods of certain "simplest" singularities which are generically stable.

Simple Lagrange Singularities. A singularity of a Lagrangian map is said to be *simple* if the set of singularities into which it can be decomposed by a small deformation of the fibration is finite (see also p. 560).

Simple singularities are classified by the Dynkin diagrams A_k, D_k, E_6, E_7, E_8 (Fig. 16.9) of the simple Lie algebras, of quivers theory, etc.; classifying also the regular polyhedra in Euclidean 3-space and the discrete subgroups of the group $\mathrm{Spin}(3) = \mathrm{SU}(2)$. See their construction in Ch. 7, pp.260-265.

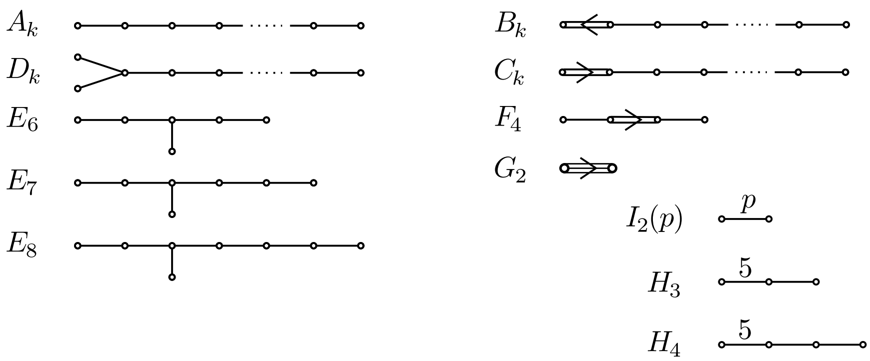


Figure 16.9: The Dynkin diagrams.

The diagrams with multiple edges, B_k, C_k, F_4, G_2, correspond to the *boundary singularities* of caustics.

Example. The focal set of a surface with boundary in Euclidean 3-space consists of three components: the focal set of the surface, the focal set of the boundary curve and the union of the normals to the surface at the points of the boundary curve (B_2 in Fig. 16.10).

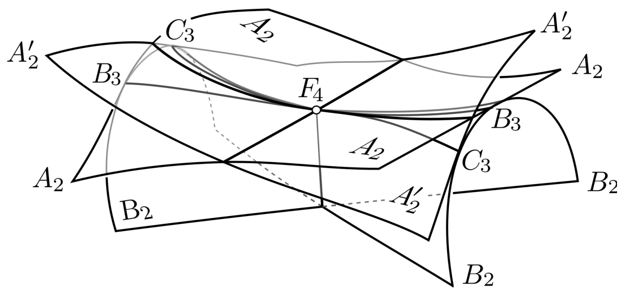


Figure 16.10: The caustic of the group F_4.

The caustic singularity F_4 (Fig. 16.10) occurs near the focal point of the surface at a point of the boundary curve where this curve is tangent to a principal curvature direction of the surface (see [111]).

The list of simple Lie algebras is a part of the list of Coxeter *Euclidean reflection groups*, most of which are *crystallographic* (Sect. 7.9). The full list of these groups, crystallographic or not, provides the classification of the first Lagrangian singularities that one has to study in symplectic geometry.

Givental Theorem ([80]). *There exists a natural bijection between the simple singularities of Lagrangian maps of Lagrangian varieties and the irreducible Coxeter groups. A variety can have a simple projection singularity only if it is* (*locally*) *a product of a smooth manifold with a curve.*

The A, D, E groups correspond to smooth curves (*and hence to the projections of smooth Lagrangian manifolds*)*. The groups with double lines in the Dynkin diagram correspond to the curves with an ordinary double point and to the boundary singularities. The remaining groups, G_2, H_3, H_4 and $I_2(q)$, correspond to the curve singularities $x^2 = y^p$ for $p = 4, 3, 3$ and $q - 2$.*

The crystalline and quasi-crystalline structures associated to Coxeter groups are crucial for studying the geometry of caustics and wave fronts in their corresponding problems of symplectic and contact geometry.

They control also the asymptotics of the integrals of the short-wave approximations and the ramification of the corresponding special functions in the complex domain. The Coxeter group represents the *monodromy group* and the crystallographic lattice is the integer homology group generated by the *vanishing cycles* on a complex hypersurface.

To explain the relations between the Coxeter groups and Lagrangian singularities one needs some basic notions of contact geometry.

16.6 Contact Geometry

Symplectic geometry of even-dimensional phase spaces has an odd-dimensional twin: Contact Geometry. The relation between contact geometry and symplectic geometry is similar to the relation between projective geometry and linear algebra. Any fact in symplectic geometry can be formulated as a contact geometry fact and vice versa.

Calculations are usually more easy in the symplectic setting and in the linear algebra of matrices and quadratic forms, but the *real understanding* of the same facts is provided by the contact and projective geometry patterns, like ellipsoids and their principal axes instead of the quadratic forms and their eigenvalues. Functions and vector fields of symplectic geometry are replaced by hypersurfaces and line fields in contact geometry.

Contact geometry is almost unknown in the mathematical physics community in spite of the fact that it provides the mathematical basis for Huygens theory of wave propagation (and hence for geometrical optics), for Gibbs thermodynamics and for optimal control theory.

The geometry of wave fronts is highly nontrivial even in the simplest case of wave propagation in Euclidean plane where the propagating fronts are equidistant curves of the initial one. Let us review a relevant recent topic.

The Inside-Out Reversal of a Front. Consider the equidistant curves of an ellipse (Fig. 16.11). As the distance from the ellipse grows the equidistant curve first shrinks and later starts to grow (if one starts from a circle it shrinks to a point). The transition from shrinking to growing is decomposed into several steps. At three steps the equidistant curve has four cusps.

This is a special case of a general theorem of symplectic and contact topology, according to which any inside-out reversal process requires at least four cusps on a moving curve at some moment. A proof for curves having no parallel co-orienting normals is given in § 16.9.2 (see [29]). The general result has been proved in 1999 by Chekanov and Pushkar [57]. Their proof depends on Sturm-Hurwitz theorem of p. 345 (see also p. 586) implying that for any function f on the circle, the function $f' + f'''$ has at least four zeros.

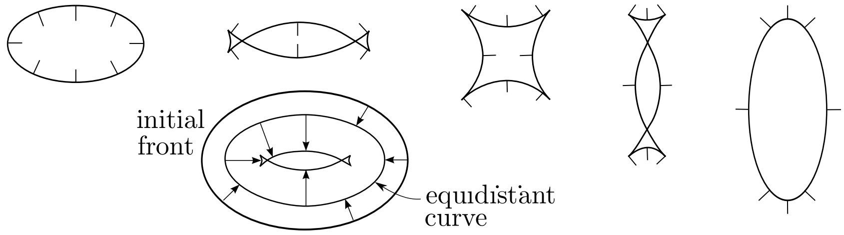


Figure 16.11: The equidistant curves of an ellipse.

This general theorem of contact topology suggests generalisations of Sturm theory to the multivalued and higher dimensional cases in the same sense that the symplectic fixed-point theorem extends Morse theory.

16.7 Contact manifolds

Contact manifolds are the odd-dimensional twins of symplectic manifolds. Each contact manifold has a symplectisation: it is a symplectic manifold whose dimension exceeds that of the given contact manifold by one. Symplectic manifolds have contactisations whose dimensions exceed their own dimensions by one.

Contact Structure. A *contact structure* on a manifold is a non degenerate field (defined below) of tangent hyperplanes called *contact hyperplanes*.

Non-degeneracy Condition. Since a field of hyperplanes is locally defined as the field of kernels of a differential 1-form, say α, the non-degeneracy condition is that the restriction of $d\alpha$ to each hyperplane $\alpha = 0$ must be a non degenerate bilinear form, that is, a linear symplectic form.

Since the symplectic spaces are even dimensional, we conclude that

Contact structures may exist only on odd-dimensional manifolds.

A *contact manifold* is a manifold provided with a contact structure.

Every 1-form α that locally determines a contact structure is called a *contact form*. In general it is defined only locally.

Remark. The non-degeneracy of the contact form α is algebraically expressed in $2n+1$-space by $\alpha \wedge (d\alpha)^n \neq 0$. It is also called "maximal nonintegrability condition".

Exercise. Verify that the non-degeneracy condition in the definition of contact structure does not depend on the choice of a special contact form but only on the field of hyperplanes. *Hint.* Such a contact form is defined up to multiplication by a function different from zero.

Example. The simplest contact manifold is Euclidean space $\mathbb{R}^3$, with coordinates (x, y, p), provided with the *standard contact structure* given by the form $\alpha = p\,dx - dy$. (Easy exercise: Verify its non-degeneracy.)

Similarly, the space $\mathbb{R}^{2n+1}$ with coordinates $(x_1, \ldots, x_n; y; p_1, \ldots, p_n)$ has the *standard contact structure* given by the form $\alpha = p\,dx - dy$.

Note that at each point the $(p_1 \ldots p_n)$-space lies in the contact hyperplane.

Problem. Consider the sphere $\mathbb{S}^{2n-1}$ of length 1 vectors in the complex Hermitian space $\mathbb{C}^n$. At each point $z \in \mathbb{S}^{2n-1}$, the Hermitian orthogonal complement to the vector iz is tangent to the sphere at z.

These orthocomplements form a field of tangent hyperplanes on the sphere $\mathbb{S}^{2n-1}$. Prove that this field of hyperplanes is a contact structure on $\mathbb{S}^{2n-1}$.

Let us define the natural isomorphisms between contact manifolds.

Contactomorphism. A *contactomorphism* between two contact manifolds M and N, is a diffeomorphism, say, from M to N, whose tangent map sends the contact structure of M to that of N. Two contact manifolds are said to be *contactomorphic*, if there exist a contactomorphism between them.

Some times one has to consider local contactomorphisms.

Darboux Theorem (on contact structures). *All contact manifolds of the same dimension are locally contactomorphic.*

That is, in a neighbourhood of every point of a contact manifold the contact structure is given in suitable coordinates by the form $p\,dx - y$.

The coordinates (x, y, p) are called *Darboux coordinates.*

Moreover, *any point of a connected contact manifold can be carried to any other by a flow of contactomorphisms.*

The classifications of complex, symplectic and contact structures on smooth manifolds is a difficult unsolved problem of modern mathematics.

16.7.1 The space of contact elements of a manifold

A *contact element* at a point of a given manifold B^n is a vector hyperplane of the tangent space of B at that point (called his *point of contact*).

All such contact elements form the *manifold E^{2n-1} of contact elements* of B, which is fibred over B. The fibre over a point $x \in B$, which consists of all contact elements at the point of contact x, is the projective space

$$\mathbb{R}\mathrm{P}^{n-1} = P(V^*)\ , \quad \text{where } V = T_x B^n\ , \quad P(V^*) = \frac{V^* \smallsetminus 0}{\mathbb{R} \smallsetminus 0}\ .$$

(The nonzero multiples of a nonzero linear form on T_xB are identified with their common kernel: a contact element at x.) The manifold E^{2n-1} is also called *projectivised cotangent bundle of* B and is denoted by PT^*B.

This manifold $E^{2n-1} = PT^*B$ possesses a single intrinsic field of hyperplanes: Writing $\pi : E^{2n-1} \to B$ for the projection along the fibres, the *tautologic hyperplane* at a point V of E^{2n-1} is the preimage of the contact element $V \subset T_xB$ under the derivative of π at our point $V \in E^{2n-1}$.

The tautologic hyperplanes form the *natural contact structure* of E^{2n-1}.

Equivalently, the hyperplanes of the natural contact structure on E^{2n-1} are defined by the *skating condition*: The velocity vector of a moving point V in E^{2n-1} belongs to the contact hyperplane if and only if the velocity of its point of contact (in B) belongs to the contact element $V \subset T_xB$.

In other words, the contact element (the skate) can rotate around the point of contact or can move tangentially to its own direction, but it can not move transversely to its direction.

Theorem 2. *The tangent hyperplanes defined by the skating condition form a contact structure on the manifold of contact elements.*

This example explains the term “contact structure”.

The Manifold of Co-Oriented Contact Elements ST^*B.
A *co-oriented contact element* of an n-dimensional manifold B is a contact element of B together with the choice of one of the two half-spaces into which it subdivides the tangent space T_xB at the point of contact x of B.

The co-oriented contact elements of B at the point of contact x form the $(n-1)$-sphere, which covers twice the above projective space,

$$\mathbb{S}^{n-1} = S(V^*) \ , \quad \text{where } V = T_xB^n \ , \quad S(V^*) = \{p \in V^* : |p| = 1\} \ .$$

The natural contact structure on the manifold ST^*B of co-oriented contact elements of B is inherited from that of PT^*B because ST^*B is a double covering of PT^*B. They are locally the same contact manifold.

16.7.2 Projective Duality

Consider the manifold of contact elements of the real projective space $\mathbb{R}\mathrm{P}^n$. In this particular space there is a nice special construction which generates such general mathematical and physical notions as the “Legendre transformation”, the “projective duality” and the “tangential coordinates”.

Consider the dual projective space $(\mathbb{R}\mathrm{P}^n)^\vee$ whose points are the projective hyperplanes of the original one. It is the projectivisation of the vector space which is dual to the vector space whose projectivisation is the original projective space,

$$\mathbb{R}\mathrm{P}^n = \frac{\mathbb{R}^{n+1} \setminus 0}{\mathbb{R} \setminus 0}\ , \qquad (\mathbb{R}\mathrm{P}^n)^\vee = \frac{(\mathbb{R}^{n+1})^* \setminus 0}{\mathbb{R} \setminus 0}\ .$$

It is clear that the projective duality is an involution, $(\mathbb{R}\mathrm{P}^n)^{\vee\vee} = \mathbb{R}\mathrm{P}^n$, because the equality $V^{**} = V$ is true for vector spaces.

The direct product of the projective space with its dual projective space

$$M^{2n} = \mathbb{R}\mathrm{P}^n \times (\mathbb{R}\mathrm{P}^n)^\vee$$

contains a remarkable smooth hypersurface $E^{2n-1} \subset M^{2n}$ which is called the *incidence hypersurface* and is formed by the pairs (x, h) such that the hyperplane h of $\mathbb{R}\mathrm{P}^n$ contains the point x of $\mathbb{R}\mathrm{P}^n$. We can also say that the hyperplane x of the dual projective space $\mathbb{R}\mathrm{P}^{n\vee}$ contains the point h of this dual space – it is just the same incidence relation.

The incidence hypersurface can be interpreted as the manifold of contact elements of the original projective space $\mathbb{R}\mathrm{P}^n$,

$$E^{2n-1} = PT^*(\mathbb{R}\mathrm{P}^n)\ ,$$

because every projective hyperplane h containing a point x defines a contact element at x (the tangent space to h at x) and is defined unambiguously by this element.

Hence, the natural projection $\pi_1 : M^{2n} \to \mathbb{R}\mathrm{P}^n$ (sending (x, h) to x) induces a fibration on E^{2n-1} over the projective space $\mathbb{R}\mathrm{P}^n$, which we have identified with the fibration of the manifold of contact elements of the original projective space. Similarly, the second natural projection, $\pi_2 : M^{2n} \to (\mathbb{R}\mathrm{P}^n)^\vee$ (sending (x, h) to h) induces a fibration on E^{2n-1}, identified with the fibration of the manifold of contact elements of the dual projective space:

$$\begin{array}{ccc} E^{2n-1} & = & PT^*(\mathbb{R}\mathrm{P}^n) \\ \downarrow & & \downarrow \\ B_1^n & = & \mathbb{R}\mathrm{P}^n\,, \end{array} \qquad \begin{array}{ccc} E^{2n-1} & = & PT^*(\mathbb{R}\mathrm{P}^n)^\vee \\ \downarrow & & \downarrow \\ B_2^n & = & (\mathbb{R}\mathrm{P}^n)^\vee\,. \end{array}$$

Each of these two interpretations of the hypersurface of incidences E^{2n-1} defines on it a contact structure which is the tautological contact structure of the manifold of the contact elements of $\mathbb{R}\mathrm{P}^n$ and that of $(\mathbb{R}\mathrm{P}^n)^\vee$.

Hence the manifold of contact elements of a projective space is the total space of *two* fibrations: The first over the original projective space and the second over the dual one – Fig. 16.12. These fibrations are the projectivised cotangent bundles $PT^*\mathbb{R}\mathrm{P}^n$ and $PT^*(\mathbb{R}\mathrm{P}^n)^\vee$. Thus the manifold of contact elements of the projective space has *two* natural contact structures.

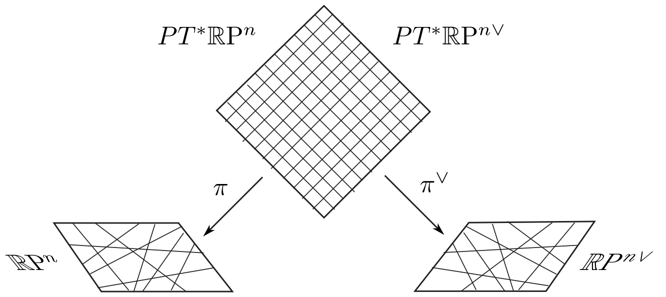


Figure 16.12: Projective duality and the pair of fibrations.

Remark. Note that the hyperplanes of $\mathbb{R}\mathrm{P}^n$ are the images of the fibres of $\pi^\vee$ under the projection π. And reciprocally, the hyperplanes of $\mathbb{R}\mathrm{P}^{n\vee}$ are the images of the fibres of π under the projection $\pi^\vee$.

The main fact of projective duality theory is the following

Theorem 3. *Both contact structures on the incidence manifold coincide.*

The published proofs are mostly computations of the expressions of both tautological structures in the natural homogeneous coordinate systems of the projective spaces; but there is a coordinate-free simple geometric proof with no calculation. We leave to the reader the pleasure to discover this geometric proof. *Hint*: Consider the product of two identical projective spaces and the involution permuting them.

Duality. Projective duality associates to a submanifold S in projective space its *dual submanifold* $S^\vee$ consisting of all hyperplanes tangent to the original submanifold S. It lives in the dual projective space.

It is classically known that the tangent line to a plane curve at an inflection point represents a semicubic cusp of the dual curve. A line tangent to the curve at two points is a selfintersection point of the dual curve (Fig. 16.13).

The affine version of projective duality is the *"Legendre transform"*. Thus contact geometry is the geometrical base of Legendre transform and of all theories based on it : thermodynamics, optimal control theory, etc. To introduce the Legendre transform, we shall first present the manifold of 1-jets.

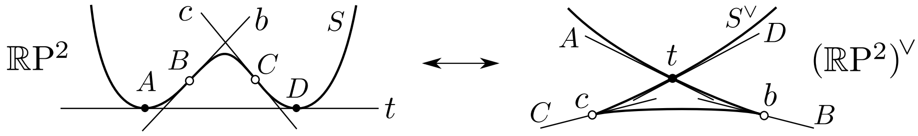


Figure 16.13: A cusp is dual to an inflection.

16.7.3 The Manifold of 1-jets of functions $J^1(M,\mathbb{R})$

Working with projective space, one has in mind both: The affine space to which one adds points at infinity, and the vector space (one dimension higher) whose 1-dimensional subspaces are the points of the projective space.

The manifold of contact elements corresponds to the second approach: it is the projectivisation (of the fibres) of the space of the cotangent bundle, and it has one dimension less than this even-dimensional phase space.

Another source of contact manifolds, corresponding to the first approach, is provided by Taylor expansions. The space of 1-jets $J^1(M,\mathbb{R}) = T^*M \times \mathbb{R}$ (defined below) has one dimension more than the phase space T^*M.

The *k-jet* of a smooth function is its Taylor polynomial of degree k:

1-jet. A 1-*jet* of a function is its Taylor polynomial of degree one. So it is a triple consisting of a point of the manifold, of the value of the function at that point and of its differential at that point: $j^1_x f = (x, f(x), d_x f)$.

Example. The manifold of 1-jets of the functions of one variable x is the 3-dimensional space considered above $J^1(\mathbb{R},\mathbb{R}) \approx \mathbb{R}^3 = \{(x,y,p)\}$.
The polynomial of degree one corresponding to the coordinates x, y, p is:

$$f(z) = y + p(z-x) \quad (+O(|z-x|^2)) \ .$$

Similarly, the space of 1-jets of the functions of n variables is the space $J^1(\mathbb{R}^n,\mathbb{R}) \approx \mathbb{R}^{2n+1} = \{(x_1,\dots,x_n,y,p_1,\dots,p_n)\}$ corresponding to

$$f(z) = y + p_1(z_1-x_1) + \dots + p_n(z_n-x_n) \quad (+\dots) \ .$$

Likewise, the space $J^1(X,\mathbb{R})$ of 1-jets of the functions on a manifold X^n can be defined using a local coordinate system on X to introduce local coordinates on the manifold $J^1(X,\mathbb{R}) = T^*X \times \mathbb{R}$, with the natural contact structure defined by the hyperplanes $\alpha = 0$ given by the contact form

$$\alpha = p\,dx - dy = p_1\,dx_1 + \dots + p_n\,dx_n - dy \,.$$

But we shall give an intrinsic (coordinate-free) definition of the contact manifold $J^1(X,\mathbb{R})$. Namely, a point $(x, f(x), d_x f)$ of $J^1(X,\mathbb{R})$ is an equivalence class of the functions whose graphs are tangent at $x \in X$ (Fig. 16.14 - this was the origin the term "jet"); its contact hyperplanes will be constructed in Theorem 4. For we consider the space $J^k(X,\mathbb{R})$ of k-jets, whose points j_x^k are defined likewise but requiring a tangency of order k of the graphs at x.

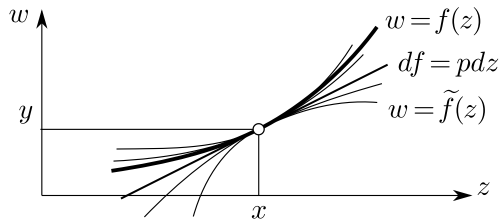


Figure 16.14: The 1-jet of a function f at a point x is the equivalence class of the functions $\widetilde{f}$ whose graphs are tangent to the graph of f at x.

Definition. The k-graph of a smooth function f on a manifold X is the submanifold of $J^k(X,\mathbb{R})$, diffeomorphic to X, formed by the Taylor polynomials $j_x^k f$ of degree k of f at all points x of X.

Example. The 0-graph is the usual graph $\{(x, f(x)) \in J^0(X,\mathbb{R}) = X \times \mathbb{R}\}$.

Theorem 4. *The tangent spaces of the* 1*-graphs of all functions on a manifold, passing through a given point of the manifold of* 1*-jets, are contained in a tangent hyperplane of the manifold of* 1*-jets at this point and fill this hyperplane densely.*

The 1-form $\alpha = p\,dx - dy$ vanishes along the 1-graph of any smooth function because $df = \sum_{k=1}^{n} (\partial f/\partial x_k)\,dx_k$. Hence in the standard local coordinates of the space of 1-jets (x_i for the argument, y for the value, p_i for the partial derivatives) the desired tangent hyperplane is given by the equation

$$dy = p_1\,dx_1 + \cdots + p_n\,dx_n\,.$$

Theorem 5. *The hyperplanes defined in Theorem* 4 *form a contact structure.*

16.7.4 Legendre Transform

We start providing a way to solve efficiently the following problem.

Problem. Find the dual curve of the plane curve S defined as the graph of a differentiable function $f:\mathbb{R}\to\mathbb{R}$, that is, $S=\{(x,y)\in\mathbb{R}^2 : y=f(x)\}$.

Since any such graph is transverse to the y-direction, we use the following *Remark. The space of* 1*-jets* $J^1(\mathbb{R},\mathbb{R})$ *parametrise all contact elements of the plane of variables* (x,y) *which are transverse to the* y*-direction* (that is, with finite slope). Namely, the contact element with slope $p_0\neq\infty$ at the point (x_0,y_0) is represented by the point (x_0,y_0,p_0) of $J^1(\mathbb{R},\mathbb{R})$ and vice versa.

Similarly, the space of 1-jets $J^1(\mathbb{R}^n,\mathbb{R})$ parametrise all contact elements of the space of variables $(x,y)\in\mathbb{R}^n\times\mathbb{R}$, transverse to the y-direction.

Affine duality. Consider the equation of the pairs of variables (x,y), $(\widetilde{x},\widetilde{y})$

$$y+\widetilde{y}=\widetilde{x}x\,. \tag{2}$$

The variables $(\widetilde{x},\widetilde{y})$ parametrise all lines transverse to the y-direction in the xy-plane, and the variables (x,y) parametrise all lines transverse to the $\widetilde{y}$-direction in the dual $\widetilde{x}\widetilde{y}$-plane.

Hence the contact element (x,y,p) corresponds to the contact element of the dual plane $(\widetilde{x},\widetilde{y},\widetilde{p})=(p,\,p\,x-y,\,x)$ and the two contact structures coincide because $\widetilde{p}\,d\widetilde{x}-d\widetilde{y}=-(p\,dx-dy)$. Moreover $(\widetilde{\widetilde{x}},\widetilde{\widetilde{y}},\widetilde{\widetilde{p}})=(x,y,p)$ and we have two projections, similar to those of the incidence manifold (Fig. 16.12), $\pi:(x,y,p)\mapsto(x,y)$ and $\pi^\vee:(x,y,p)\mapsto(\widetilde{x},\widetilde{y})=(p,p\,x-y)$.

We get the following parametrisation of the dual curve $S^\vee$:

$$x\mapsto(\widetilde{x},\widetilde{y})=(f'(x),f'(x)x-f(x)), \tag{3}$$

which is the graph of the (multivalued) function $\widetilde{f}(\widetilde{x})=\widetilde{x}\,x-f(x)$ called the *Legendre Transform* of the function f (see Fig. 16.15).

Exercise. Parametrise and draw the dual curve of the plane curve defined by the equation $y=x^4-tx^2$, for $t=1,0,-1$. Draw the union of dual curves in space-time for $t\in[-1,1]$.

Problem. Find the dual curve of the curve given by the equation $y=x^\alpha/\alpha$.

Answer. The equation of the dual curve is $\widetilde{y}=\widetilde{x}^\beta/\beta$, where $1/\alpha+1/\beta=1$.

Higher dimensions. Given a hypersurface of the affine space, defined as the graph of the function $y=f(x_1,\dots,x_{n-1})$, its dual hypersurface is the "graph of a multivalued function" $\widetilde{f}$ called *the Legendre transform* of the function f, $\widetilde{y}=\widetilde{f}(\widetilde{x}_1,\dots,\widetilde{x}_{n-1})$. Its coordinate definition is as follows.

The new independent variables $\widetilde{x}_k$ are the $n-1$ "slopes" p_k : $\widetilde{x}_k = p_k := \partial f / \partial x_k, \ k = 1, \dots, n-1,$ and the new function is given by

$$\widetilde{f}(\widetilde{x}) = \sum \widetilde{x}_k \, x_k - f(x)\,,$$

where the $x_k = x_{k*}(\widetilde{x})$ are determined by the $n-1$ "slopes" (Fig. 16.15).

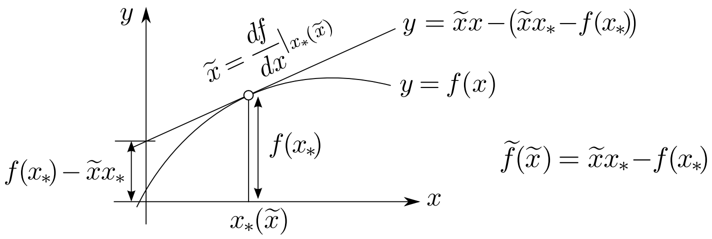


Figure 16.15: Legendre transform of the function f of the variable x to the function $\widetilde{f}$ of the dual variable $\widetilde{x}$.

Exercise. Find the Legendre transform of the quadratic form $f(x) = \sum m_k \dfrac{x_k^2}{2}$.

Answer. $\widetilde{f}(\widetilde{x}) = \sum \dfrac{\widetilde{x}_k^2}{2m_k}$ (this equality is useful to study confocal quadrics - Ch. 17).

Corollary. *For the quadratic forms* $\widetilde{f}(\widetilde{x}) = f(x)$.

Remark. In physics the quadratic form f is called "kinetic energy", m_k are the masses, and x_k are the velocities. The new variables, $\widetilde{x}_k = m_k x_k$, are called the *kinetic moments*, and $\widetilde{f}(\widetilde{x}) = f(x)$ is still the kinetic energy.

16.8 Legendrian Submanifolds

Legendrian submanifolds of contact manifolds are the twins of the Lagrangian submanifolds of symplectic manifolds. One needs Legendrian submanifolds, for example, to explain the properties of dual hypersurfaces, Legendre transform and propagation of wave fronts.

Definition. A submanifold of a contact manifold is said to be *integral* if its tangent space at every point belongs to the contact hyperplane. A *Legendre submanifold* of a contact manifold is an integral submanifold of maximal dimension: equal to $n-1$ for a $(2n-1)$-dimensional contact manifold.

The "contact Weinstein principle" should say that *in contact geometry every interesting object is a Legendrian submanifold*.

Problem. Prove that the contact manifold of the contact elements of a smooth base manifold of dimension n has no integral submanifolds of the contact structure, whose dimension exceeds $n-1$.

Legendrian lift. Given a manifold B^n, the contact elements tangent to any submanifold $S \subset B$ of positive codimension form a Legendrian submanifold L_S of the manifold $E^{2n-1} = PT^*B$ of contact elements of B. This submanifold L_S is called the *Legendrian lift* of S in E^{2n-1}. For example, the fibre of $\pi : E^{2n-1} \to B^n$ over a point of B (that is, the manifold of all contact elements tangent to that point) is the Legendrian lift of that point.

1-Graphs. It is evident from Theorems 4 and 5 that the 1-graph of any smooth function on a manifold X is a Legendrian submanifold of the space of 1-jets $J^1(X^n, \mathbb{R})$. We can see it also by observing that the 1-graph of a function is the Legendrian lift in $J^1(X, \mathbb{R})$ of the graph of that function (which consists of the contact elements of $X \times \mathbb{R}$ tangent to that graph).

Moreover, each fibre of the "forgeting derivatives projection"

$$J^1(X, \mathbb{R}) = T^*X \times \mathbb{R} \to X \times \mathbb{R}\,, \qquad (x, p, y) \mapsto (x, y)\,,$$

is a Legendrian submanifold because the tangent space to the fibre is contained in the contact hyperplane $\sum p_i\, dx_i = dy$ (note that $\sum p_i\, dx_i$ is the action 1-form on T^*X; and the fibration $J^1X \to X \times \mathbb{R}$ is given by the projection along the (Lagrangian) fibres of the cotangent bundle of X).

Quasi-functions. Since any non vertical Legendrian submanifold of the manifold of 1-jets $J^1(X, \mathbb{R})$ is (at least locally) the 1-graph of a function on X, it is natural and productive to consider the general Legendrian submanifolds of $J^1(X, \mathbb{R})$ as generalised (in general multivalued) functions. For example, one can extend Morse inequality to such "quasi-functions" and prove generalised "symplectic fixed point theorems" (see p. 628).

16.9 Legendrian fibrations, maps and fronts

The manifold of contact elements and the space of 1-jets of functions together with their natural fibrations $PT^*B \to B$ and $J^1(M, \mathbb{R}) \to J^0(M, \mathbb{R})$, whose fibres are Legendrian, play an important role in physics and mathematics. Their applications have lead to consider the following **definitions**:

• A *Legendrian fibration* is a fibration whose fibres are Legendrian.
• A *Legendrian map* is the projection of a Legendrian submanifold of the space of a Legendrian fibration to its base along its Legendrian fibres.
• The image of a Legendrian map is called its *front*.
(as a rule the front is a hypersurface in the base space of the Legendrian fibration; like a caustic is, as a rule, a hypersurface of the base space of its Lagrangian fibration.)

Example. Both fibrations $\pi : E^{2n-1} \to \mathbb{R}\mathrm{P}^n$ and $\pi^\vee : E^{2n-1} \to \mathbb{R}\mathrm{P}^{n\vee}$ shown in Fig. 16.12 are Legendrian (here, $PT^*\mathbb{R}\mathrm{P}^n = E^{2n-1} = PT^*\mathbb{R}\mathrm{P}^{n\vee}$).

Tangential map - dual fronts. Given a smooth submanifold S of the projective space $B = \mathbb{R}\mathrm{P}^n$, consider its Legendre lift L_S in the manifold $E^{2n-1} = PT^*(\mathbb{R}\mathrm{P}^n)$ of contact elements (Fig. 16.16). Interpreting E^{2n-1} as the manifold of contact elements of the dual projective space and projecting $L_S \subset E^{2n-1}$ to the second base space along the fibres of $\pi^\vee$, we get a Legendrian map called *tangential map** of S, $\tau_S = \pi^\vee_{|L_S} : L_S \to (\mathbb{R}\mathrm{P}^n)^\vee$.

Therefore $S^\vee$, the dual submanifold of S, is the front of the tangential map τ_S and, clearly, the contact elements of $\mathbb{R}\mathrm{P}^{n\vee}$ tangent to $S^\vee$ form the smooth Legendrian submanifold L_S, that is, $L_{S^\vee} = L_S$.

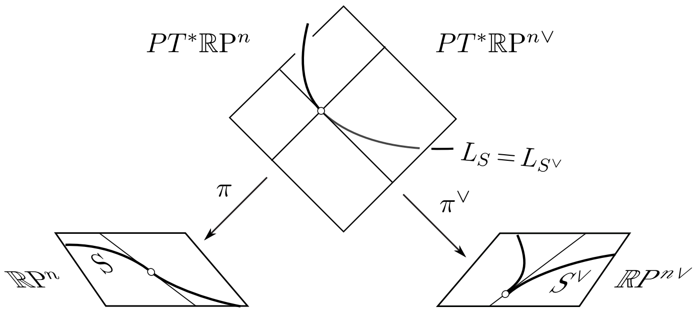


Figure 16.16: The dual hypersurface $S^\vee$ of a hypersurface S in projective space.

Corollary. *The graph of the (multivalued) Legendre transform of a smooth function is the front of a Legendrian map.*

It follows with no calculation (Theorem 3, p. 605) that *the dual of the dual hypersurface is the initial hypersurface*, $S^{\vee\vee} = S$, at least if their (common) Legendre lift is in general position with respect both Legendre fibrations.

*The reader should not confuse the tangential map of a hypersurface with the tangent map of a smooth map, defined in p. 29.

Hint: any Legendrian submanifold of the space E of contact elements of a manifold B which is transverse to the fibres of the natural fibration $E \to B$ is (at least locally) the manifold of tangent hyperplanes of a smooth hypersurface in the base manifold B.

Therefore the Legendre transform is an involution: $\widetilde{\widetilde{f}} = f$.

Problem. Find the dual curve of the plane curve S defined by the equation $y = x^3$.

Solution. Its Legendrian lift L_S (the 1-graph of $f(x) = x^3$) has the parametrisation $x = x$, $y = x^3$, $p = 3x^2$. Thus the dual curve $S^\vee$ is the semicubic parabola parametrised by $\widetilde{x} = 3x^2$, $\widetilde{y} = 3x^2 \cdot x - x^3 = 2x^3$ (see Fig. 16.16) and the Legendre transform is

$$\widetilde{y} = c\widetilde{x}^{3/2} \qquad (c = 2/(3\sqrt{3}))\ .$$

Inflection $\leftrightarrow$ Cusp. Given a plane curve S, the points where its Legendre lift L_S is tangent to a fibre of the fibration $\pi^\vee$ correspond to the inflections of S (because at these points the curve S has 3-point contact with its tangent line, and the Legendre lifts of the straight lines are the fibres of $\pi^\vee$). Whence the semicubic cusp of $S^\vee$ in this example.

In general, the singularities of a front occur at points for which the tangent space of its Legendre lift is not transverse (in the contact hyperplane) to the tangent space of the fibre.

Duals of Generic Surfaces in $\mathbb{RP}^3$. For the 19th century mathematicians the following facts on smooth surfaces in general position were well known:

1. The elliptic and hyperbolic domains of a surface S (see p.318) correspond to the respective elliptic and hyperbolic domains of its dual surface $S^\vee$;
2. The parabolic curve of S corresponds to the cuspidal edge of its dual $S^\vee$;
3. Godrons of S correspond to swallowtail points of its dual $S^\vee$ (Fig. 16.17);

Korteweg called *conodal* the curve formed by the points (the dotted line in Fig. 16.17) that correspond to the self-intersection line of the dual front, which is important in thermodynamics. Two points of a surface are called *conodes* if there is a plane tangent to the surface at both points. Hence:

The closure of the locus of the connodes of a surface is the conodal curve.

4. At a godron the conodal curve is tangent to the parabolic curve and consists of elliptic (hyperbolic) points near a positive (negative) godron.

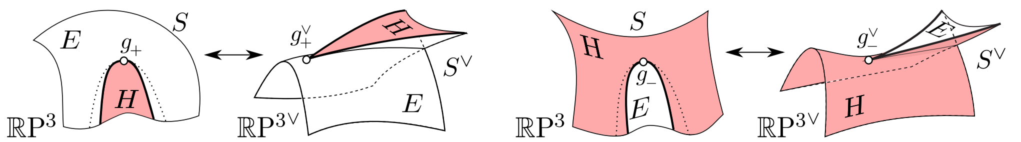


Figure 16.17: Duality godron $\leftrightarrow$ swallowtail for the two types of godrons.

Exercise. Verify item 4 (depicted in Fig. 16.17). *Hint*: Use Platonova's normal form (2) of p.318 and apply the Legendre transform to parametrise the dual front.

The Legendrian nature of the tangential map explains these properties.

16.9.1 Legendrian Gibbs' Thermodynamics

Daily experience shows us that materials can evaporate, freeze, melt and acquire various physical states. Thermodynamics studies the evolution and properties of macroscopic systems in terms of pressure, temperature, etc.

According to Gibbs, the laws of thermodynamics are defined by a contact structure on the 5-space of thermodynamical states $\{(S, V, U, T, p)\}$, where the variables S (entropy), V (volume), U (internal energy), T (temperature) and p (pressure) satisfy the equations $\partial U/\partial S = T$ and $\partial U/\partial V = -p$ along the equilibrium states of each substance. GIBBS' DISCOVERY was that:

The contact hyperplanes are given by the differential relation

$$dU = TdS - p\,dV$$

and each substance is represented as a Legendre 2*-surface in this* 5*-space.*

Different physical states of the same substance correspond to different points of its Legendre surface (and the states evolve inside this surface).

All the thermodynamic properties of a system can be found from any of the four energies (Thermodynamic Potentials): Internal energy U, Gibbs free energy G, Enthalpy H and Helmholtz free energy F. Depending on the type of process, one of them provides the most convenient description.

Four energy functions (thermodynamic potentials) are used in thermodynamics of chemical reactions and non-cyclic processes. They are the internal energy (U), the enthalpy (H), the Helmholtz free energy (F) and the Gibbs free energy (G).

Depending on the type of non-cyclic process, one of the four energies (Thermodynamic Potentials) provides the most convenient description of all thermodynamical properties of the given system: Internal energy U, Gibbs free energy G, Enthalpy H and Helmholtz free energy F.

In fact the graphs of these energy functions are fronts of the Legendre submanifold of the substance, projected along four Legendre fibrations.

Phase change vapour-liquid. To study a one-component fluid, we project its Legendre surface L along a Legendrian fibration, getting a front in which we follow the fluid evolution. One often takes, as such front, the graph Γ of the internal energy $U = U(S, V)$ (L is clearly the 1-graph of U) or its dual surface: The graph $\Gamma^\vee$ of the *Gibbs' energy* $G = U - TS + PV$ (G is the Legendre transform of $-U$) of which each tangent plane represents a thermodynamical state (see an isothermal ($T =$ const.) in Fig.16.18-left).

Among the godrons of the smooth surface Γ only the positive ones (like in Fig. 16.17-left) represent stable equilibria. The "metastable" equilibrium states are the elliptic points between the parabolic and conodal curves; the hyperbolic domain represents the unstable equilibrium states; and the stable equilibrium states are the elliptic points that are separated from the parabolic curve by the conodal curve.

Two points of Γ with the same (bi)tangent plane are called *connodes.* They represent two distinct states with the same pressure, temperature and Gibbs' energy; the segment they define is called *tie line* (of these states).

Consider a gaseous fluid in stable equilibrium (Fig. 16.18, state a). If we increase the pressure p, keeping T = const., the state will move inside the elliptic domain of stable states (of Γ). When the state (approaches and) reaches the conodal curve (state b), the system lows its energy U going along the tie line (instead of going on the surface) and splitting into the two phases represented by the connodes (b and e) of the tie line : gas and liquid; their respective proportions vary continuously from 1 to 0 and from 0 to 1.

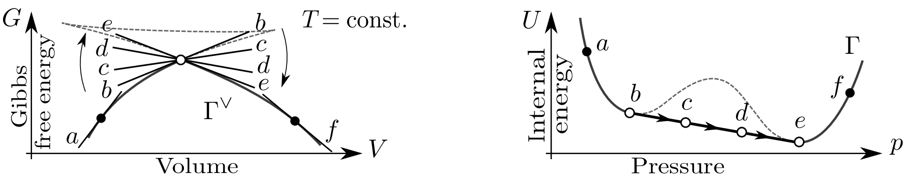


Figure 16.18: A phase transition gas-liquid, described in an isothermal section T = const. of the graph of the Gibbs energy and in its dual.

As the fixed temperature T is higher, the godron is approached and the tie line becomes smaller, implying that the coexisting liquid and gas phases become more similar : the liquid phase expands (lowering its density), while the gas phase reduces (growing its density). At last, there is a *critical temperature* in which the tie line shrinks to a godron and the two phases become identical. Over the critical temperature, there is only one phase, called *supercritical fluid* (with no distinction between liquid and gas).

Thus the graph Γ of the energy U of a real fluid contains the "piece of ruled surface" formed by the tie lines joining the coexisting phases.

16.9.2 The Cusps for Reversing Inside-Out a Hedgehog

We shall construct the natural contactomorphism $J^1(\mathbb{S}^{n-1}, \mathbb{R}) \simeq ST^*\mathbb{R}^n$ together with the "support function". Then we shall use them to prove "the reversal front four-cusps theorem" and the classical four-vertex theorem.

Problem. Consider the manifold $J^1(\mathbb{S}^{n-1}, \mathbb{R})$ of the 1-jets of functions on the unit sphere $\mathbb{S}^{n-1} = \{x : |x| = 1\}$ in Euclidean space $\mathbb{R}^n$.

Prove that this contact manifold is naturally contactomorphic to the contact manifold of the co-oriented contact elements of Euclidean space $\mathbb{R}^n$.

Solution. We shall construct a natural contactomorphism that sends the natural contact structure of the manifold of 1-jets onto the natural contact structure of the manifold of the co-oriented contact elements of $\mathbb{R}^n$.

A co-oriented contact element in Euclidean space $\mathbb{R}^n$ is determined by its point of contact $Q \in \mathbb{R}^n$ and by its co-orienting unit normal vector $P \in \mathbb{S}^{n-1}$, directed to the chosen co-orienting half-space of $T_Q\mathbb{R}^n$. The natural contactomorphism sends the co-oriented contact element (Q, P) to the 1-jet (q, y, p) of a function on $\mathbb{S}^{n-1}$ (at a point $q \in \mathbb{S}^{n-1}$, with value $y \in \mathbb{R}$, and differential $p \in T_q^*\mathbb{S}^{n-1}$), given by the simple formulas:

$$q = P\ , \quad y = \langle\, Q\ ,\ P\,\rangle\ , \quad p = Q - yP\ . \tag{4}$$

The inverse diffeomorphism is constructed in Fig. 16.19: $P = q\,,\ \ Q = yq{+}p\,.$

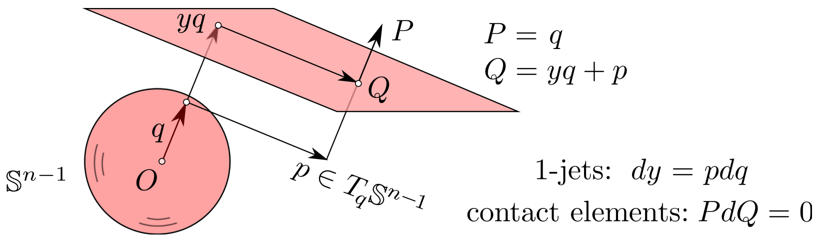


Figure 16.19: Natural contactomorphism from the manifold of 1-jets of functions on the sphere to the manifold of cooriented contact elements of Euclidean space.

Support Function. To prove that this construction provides a contactomorphism, consider a smooth co-oriented hypersurface X in $\mathbb{R}^n$, and a contact element (Q, P) tangent to it, that is, Q is a point of X and P is the co-orienting unit vector normal to X at Q. The tangent $(n-1)$-subspace to this hypersurface at the point Q intersects the orthogonal ray, starting at the origin, at some point yP. The "*support function*" of our hypersurface X is the function on the sphere $\mathbb{S}^{n-1}$ whose value y at the point $q = P$ is the oriented (algebraic) distance from O to the tangent plane of the hypersurface at the point Q (at which the oriented normal is $q = P$), that is, $y = \langle Q\,,\, P\rangle$.

This construction sends the Legendre submanifold (of $ST^*\mathbb{R}^n$) formed by the contact elements tangent to the hypersurface $X \in \mathbb{R}^n$ to the Legendre submanifold (of $J^1(\mathbb{S}^{n-1}, \mathbb{R})$) formed by the 1-jets of the support function of that hypersurface, implying the correspondence of both natural contact structures, transformed one to the other by our map (4).

These contactomorphisms make it possible to apply the geometry of the spaces of contact elements to the study of Taylor series and vice versa.

The support function is very useful despite that it is well defined modulo the choice of the origin O (being the base of Convex Body Theory).

Hedgehogs. For "*hedgehog hypersurfaces*" (i.e., without parallel co-orienting normals at distinct points, like the curves of Fig. 16.11) the support function is a genuine function, while for hypersurfaces with parallel co-orienting normals at different points, the "function" becomes multivalued.

Remark. To study plane curves, it is useful to consider Euclidean plane as the plane of complex numbers. Since unit vectors are written as $e^{i\vartheta} \in \mathbb{S}^1$ and the inner product of two vectors $v = v_1 + iv_2$, $w = w_1 + iw_2$ equals $\langle v, w\rangle = \mathrm{Re}(v\overline{w})$, the support function of a curve $Q = Q(\vartheta)$ (parametrised by its normal vector $P = e^{i\vartheta}$) is given by

$$h(\vartheta) = \mathrm{Re}(Q(\vartheta)e^{-i\vartheta})\,.$$

Exercise. Compute the support function of the standard astroid. It is parametrised by $Q(\vartheta) = 3(\cos\vartheta, -\sin\vartheta) - (\cos 3\vartheta, \sin 3\vartheta)$.

Solution. In complex notation $Q(\vartheta) = 3e^{-i\vartheta} - e^{3i\vartheta}$. Computing $Q'(\vartheta)$ one gets that $e^{i\vartheta} \in \mathbb{S}^1$ is a normal vector to the astroid at $Q(\vartheta)$. Hence $h(\vartheta) = \mathrm{Re}(Q(\vartheta)e^{-i\vartheta}) = 2\cos 2\vartheta$.

To recover a hypersurface from its support function $y(q)$, one uses the inverse of (4) to get the parametrisation $\mathbb{S}^{n-1} \to \mathbb{R}^n$, $q \mapsto Q(q) = y(q)q + p$, which is the inverse of the Gauss map (Fig. 16.19).

In the case of a plane curve, we write $q = e^{i\vartheta} \in \mathbb{S}^1$, $y = h(\vartheta)$ and $p = h'(\vartheta)ie^{i\vartheta}$ ($ie^{i\vartheta}$ is the orienting unit vector of $T_q\mathbb{S}^1$) so that

$$Q(\vartheta) = h(\vartheta)e^{i\vartheta} + ih'(\vartheta)e^{i\vartheta}\,. \tag{5}$$

Exercise. Draw the curve whose support function is $h(\vartheta) = 2\cos k\vartheta$ for $k = 3, 4, 5$.

Exercise (Propagation of equidistant wave fronts). Prove that *if h is the support function of a hedgehog hypersurface in Euclidean space, then, for every $t \in \mathbb{R}$ the support function of the t-equidistant hypersurface is the function $h + t$.*

Example. The support function of the equidistant circles of centre (a, b), with "signed radius" t, is $h_t(\vartheta) = a\cos\vartheta + b\sin\vartheta + t$, $t \in \mathbb{R}$ (different signs of t correspond to opposite co-orientations of the circle). Indeed, writing $Q_t(\vartheta) = (a + t\cos\vartheta, b + t\sin\vartheta)$ we have $P(\vartheta) = (\cos\vartheta, \sin\vartheta)$ and hence $h_t(\vartheta) = \langle Q_t(\vartheta), P(\vartheta)\rangle = a\cos\vartheta + b\sin\vartheta + t$.
Remark. The average value of h equals the signed radius t of the cooriented circle.

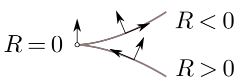
Proposition. *The signed radius of curvature of a plane cooriented curve Q with support function h is given by $R(\vartheta) = h''(\vartheta) + h(\vartheta)$.*

Proof. As we vary ϑ, the normal line at $Q(\vartheta)$, $\ell(\vartheta, \lambda) = Q(\vartheta) - \lambda e^{i\vartheta}$, moves rigidly. Since its instantaneous centre of rotation is the centre of curvature at $Q(\vartheta)$, the signed radius of curvature $R(\vartheta)$ is the value $\lambda = \lambda(\vartheta)$ given by the condition $\partial\ell/\partial\vartheta = 0$. Using (5), we get $\lambda(\vartheta) = h''(\vartheta) + h(\vartheta)$. □

Proof of 4-Vertex Theorem. By the Proposition, $R'(\vartheta) = h'''(\vartheta) + h'(\vartheta)$. The condition $k \geqslant 2$ of Sturm-Hurwitz theorem is satisfied only by functions representable as $f = g''' + g'$ (for some periodic "potential" g) because the operator $\partial(\partial^2 + 1)$ kills the constants and the first order terms of the Fourier series ($\cos'''\vartheta + \cos'\vartheta = 0$, etc.). Hence $R'(\vartheta)$ has at least four zeros. □

Remark (Cusps). Since at the cusps the velocity vanishes (and the unit tangent vector reverses its orientation), they correspond to the zeros of $h'' + h$. Moreover, at a cusp, the centre of curvature is the cusp itself:

$$Q'(\vartheta) = (h''(\vartheta) + h(\vartheta))ie^{i\vartheta}$$



.

Theorem 6 (Hedgehog Reversal Four Cusp Theorem [30])**.** *A generic path connecting a circle with a circle with reversed coorientation in Euclidean plane in the space of hedgehogs contains a front with at least four cusps.*

Proof. Since the average values of the support functions of the initial and final circles have opposite signs (being equal to their respective signed radii), there exists an intermediate front in the path for which h (and hence $h'' + h$) is orthogonal to the constant function (i.e. the average value of h is zero).

Sturm-Hurwitz Theorem implies that for this front the function $h'' + h$, as a function in ϑ (which is always orthogonal to $\cos\vartheta$ and to $\sin\vartheta$), has at least four zeros. Therefore the front has at least four cusps. □

This theorem belongs to contact topology. In fact, our restriction to the space of hedgehogs (fronts without inflections) is a strong way to avoid co-oriented self-tangencies of the fronts along the front-reversal path, that is, to avoid self-intersections of their Legendre lifts (which are embedded knots). The general formulation of the theorem (conjectured in [30]), which only requires to avoid self-intersections of the Legendre knots of the fronts, was proved in [57] by Chekanov and Pushkar using Sturm-Hurwitz theorem.

J. Sturm (1836) proved his theorem for the "algebraic" case of *trigonometric polynomials* (i.e., Fourier series with a finite number of terms). Hurwitz (1903) extended it to the case of smooth periodic functions. The importance of Sturm-Hurwitz theorem as well as of its extensions in contact geometry were discovered in [30] (see also [122, 83, 31]).

16.9.3 Moving Wave Fronts and Huygens Principle

The term *front* originates from the following example. Consider a hypersurface in a Euclidean or Riemannian space. Suppose that the hypersurface is *co-oriented*, i.e., equipped with a field of normal vectors.

The *t-equidistant hypersurface* of the given hypersurface is the set of endpoints of the co-orienting normal vectors of length t. In the Riemannian case one takes the endpoints of the geodesic segments of length t normal to the hypersurface at their initial points. One may think of the propagation of some perturbation (light, sound, epidemic) with velocity one. If the initial perturbation was bounded by an initial hypersurface, the front of its propagation at time t will be bounded by the t-equidistant hypersurface (at least for the case of the propagation from a convex body in Euclidean space).

Huygens Theorem 1. *The equidistant hypersurfaces are fronts of Legendrian maps.*

To prove it, we have to consider the *geodesic flow of co-oriented contact elements*: It is the one-parameter group of diffeomorphisms g^t of the manifold of co-oriented contact elements of the given Euclidean or Riemannian manifold which send each contact element normal to a geodesic to the contact element normal to the same geodesic at the point situated at distance t from the original point (along the geodesic) in the co-orienting direction.

Huygens Theorem 2. *The geodesic flow of the co-oriented contact elements preserves the natural contact structure of the manifold of the co-oriented contact elements.*

This deep result of contact geometry has been rediscovered many times, although expressed in other terms, under different names in calculus of variations, optimal control, and nonlinear programming theories. We therefore leave the pleasure of finding a geometric proof to the reader.

The first Huygens theorem follows, since the contactomorphism g^t of the flow send the Legendrian submanifold of contact elements tangent to the initial hypersurface to the contact elements "tangent" to its t-equidistant.

Singularities. Generic propagating wave fronts unavoidably acquire singularities. In practical problems (for example, in oil exploration using seismic waves) one has to be aware of all front singularities that generically one can meet. These singularities are the same as those of dual hypersurfaces or those of graphs of multivalued Legendre transforms of generic smooth functions:

Singularities of Fronts in Surfaces. Their only generic singularities are the semicubic cusp (A_2) and the transverse intersections of two branches of the front (A_1^2) Fig. 16.20. One see them on the equidistants of an ellipse (Fig. 16.11) and on the dual of a generic smooth plane curve (Fig. 16.13).

Singularities of Fronts in 3-Space. Their only generic singularities are the semicubic cuspidal edge (A_2), the *swallowtail surface* (A_3) and the transverse intersections of two or three branches (A_1^2), (A_1A_2), (A_1^3) Fig. 16.20.

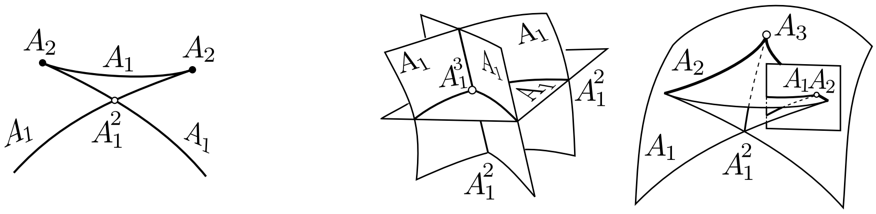


Figure 16.20: The generic singularities of a front in 2-space and in 3-space.

These short lists follow from the classification of Legendre singularities:

16.10 Classification of Front Singularities

We start pointing out that every Legendrian fibration is "locally equivalent" to the projectivised cotangent bundle $PT^*B \to B$. Namely,

Theorem 7. *All Legendrian fibrations of contact manifolds of the same dimension are locally isomorphic: There exists a local contactomorphism of the fibred spaces which takes fibres into fibres.*

Proof. It suffices to construct a local isomorphism of an arbitrary Legendrian fibration to the manifold of contact elements of some base manifold.

Given a point of the total space of a Legendre fibration, the projection of its contact hyperplane along the fibres, into the base, is a contact element of the base space whose point of contact is the projection of the fibre.

This defines a map from the total space of an arbitrary Legendre fibration to the space of contact elements of its base (i.e. to the projectivised cotangent bundle of the base) which clearly transforms fibres into fibres. This map is a (local) contactomorphism because it sends a moving point inside a fibre to a contact element of the base which turns around the point of contact with nonzero velocity (since the fibres are integral submanifolds of maximal dimension). This matches with the skating condition defining the natural contact structure of the space of contact elements of the base. □

Incidentally, we have proved that the fibres have a (local) projective structure intrinsically defined by the contact structure and the fibration.

Almost any Legendrian map is locally contactomorphic to the tangential map of a hypersurface. The exceptions are infinitely degenerate (see below)

Example. A *Legendrian collapse* is a Legendrian map that sends the Legendrian submanifold of the contact elements tangent to a submanifold that is not a hypersurface (it could be a point) to the base. Its image is the original submanifold. Under a small generic perturbation of the Legendrian manifold this exceptional Legendrian map becomes generic. Fig. 16.11 shows what happens to a point front in the plane under some generic perturbations.

The *Legendrian singularities* are the singularities of Legendrian maps.

16.10.1 Generating Hypersurfaces of Legendre maps

The classification of Legendre singularities is reducible to the study of families of hypersurfaces in the same way as the study of Lagrangian singularities has been reduced to the classification of families of functions.

A Lagrangian submanifold in the space of the cotangent bundle of some base manifold can be lifted (at least locally) to a Legendre submanifold of the manifold of 1-jets of functions on the base manifold: $z = \int_{q_0}^{q} p\, dq$.

This Legendre submanifold projects to the initial Lagrangian submanifold by the natural projection $J^1(M, \mathbb{R}) \twoheadrightarrow T^*M$ ('forgetting function values').

Thus, the generating family of a Lagrangian submanifold defines a Legendre submanifold in the space of 1-jets of functions of the variables q:

$$L = \{(q, p\,; z) : \exists x : F_x = 0, p = F_q, z = F(x, q)\}.$$

The projection $(q, p\,; z) \mapsto (q\,; z)$, being a Legendre fibration (for the contact structure $dz = p\,dq$), defines a Legendre map of the Legendre submanifold L onto its front (which lives in the space of coordinates $(q\,; z)$).

Example. The generating family of A_2, $F(x, q) = x^3 + qx$, defines the Legendre smooth curve in 3-space $\{(q, p\,; z)\}$ (Fig. 16.21):

$$q = -3x^2\,,\ p = x\,,\ z = -2x^3$$

His front (that is, his projection to the $(q\,; z)$-plane) has a semicubic cusp.

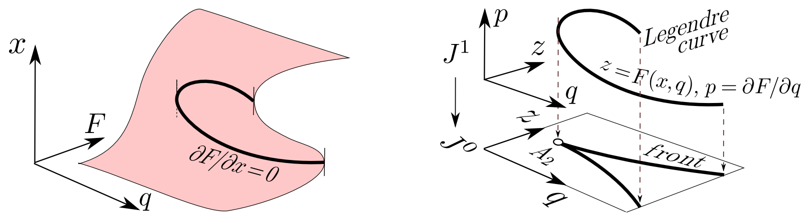


Figure 16.21: Generating family of the Legendre singularity A_2 and its front.

The generic projections of curves to the plane have no cusp. But here the cusp is stable (unremovable) under small perturbations of the Legendre curve, provided that the perturbed curve remains Legendrian.

Exercise. Verify that $F(x, q) = x^4 + q_1 x^2 + q_2 x$, the generating family of A_3 (see Sect. 16.5), defines a front diffeomorphic to the swallowtail surface.

Theorem 8. *The fronts defined by the generating families* A_k, D_k, E_k (on p. 597) *are locally diffeomorphic* (*in the complex domain*) *to the varieties of nonregular orbits of the Weyl reflection groups that correspond to the simple Lie groups* A_k, D_k, E_k (this correspondence is explained in § 7.9).

16.10.2 Relation between fronts and reflection groups

A reflection group in $\mathbb{R}^n$ acts also on $\mathbb{C}^n$. *The space of orbits of this complex action is a manifold* M *diffeomorphic to* $\mathbb{C}^n$ (see p. 138 and Vieta map below).

Discriminant. The *discriminant* of a Euclidean reflection group is the hypersurface of singular orbits in the manifold of orbits of the group.

Example. Consider the group A_2 generated by the reflections of the plane in three symmetrical mirrors (Fig. 16.22). A generic orbit consists of six points.

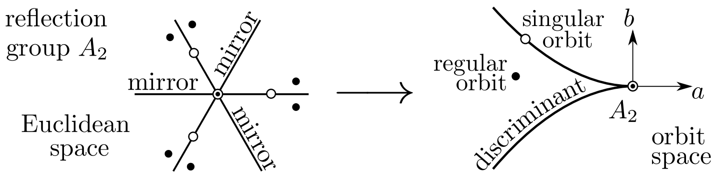


Figure 16.22: The Vieta map and the discriminant of the reflection group A_2.

It is convenient to take as our plane the plane $x_1+x_2+x_3=0$ in 3-space. The group is then generated by the transpositions of any two coordinates (that is, reflections in the planes $x_i = x_j$). It is the group of permutations of the three coordinates. Since an orbit is specified by an unordered triple of coordinates, the manifold of orbits is the set of polynomials x^3+ax+b, whose roots are those coordinates. The singular orbits consist of fewer elements than the ordinary ones, being the orbits of the points on the mirrors. The corresponding polynomials are those that have double roots.

The polynomials with double roots form the curve given by the equation $4a^3 + 27b^2 = 0$ on the plane (a, b) of cubic polynomials $x^3 + ax + b$ (prove it!). It is the discriminant of the reflection group A_2 (Figs. 16.22 and 5.8).

Example. Similarly to A_2, the action of the group A_n (of permutations of the coordinates $x_1, \ldots, x_{n+1}$ in $\mathbb{C}^{n+1}$) is reducible to his action on the space

$$\mathbb{C}^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{C}^{n+1} : x_1 + \cdots + x_{n+1} = 0\}\,.$$

The orbit space is the manifold of polynomials $x^{n+1} + a_2x^{n-1} + \cdots + a_{n+1}$. The Vieta map sends the set of roots $(x_1, \ldots, x_{n+1})$ to the set of coefficients $(a_2, \ldots, a_n)$ by means of the elementary symmetric functions of the roots $a_i = \sigma_i(x_1, \ldots, x_{n+1})$ (here $a_1 = \sigma_1(x_1, \ldots, x_{n+1}) = x_1 + \cdots + x_{n+1} = 0$).

The discriminant of A_n is the set of polynomials $x^{n+1}+a_2x^{n-1}+\cdots+a_{n+1}$ having at least one multiple root (the "generalised swallowtail", which we had studied in § 5.1.1 to describe the braid group $\mathrm{Br}(n+1)$).

Theorem 9. *The discriminant of a Euclidean reflection group of type A, D or E is the front of a generic Legendrian map at a simple singularity.*
All fronts of simple generic Legendrian singularities are locally diffeomorphic to these discriminants (or to unions of transverse branches of them).

These reflection groups are classified in terms of the mirrors of their generating reflections. The description of these sets of mirrors by their Dynkin diagrams is given in pp. 260-265 (see also Fig. 16.9 in p. 598).

The Legendre singularities of types A, D, E are the only stable and simple singularities (for Legendre maps of manifolds of any dimension).

Theorem 10. *The generic fronts of dimension $n \leqslant 5$ have only singularities Legendre equivalent to A_k, D_k, E_k, $k \leqslant n+1$.*

In higher dimensions occur generic singularities, besides the A, D, E types, having *moduli* (continuous parameters that classify part of the objects).

16.10.3 Caustics of Reflection Groups

Having associated a front to any Euclidean reflection group, one can also associate to it a caustic.

Example (**Caustic singularity** A_3). The manifold formed by the orbits of the reflection group A_3 (which is the group of symmetries of a tetrahedron) is the space of polynomials $x^4 + ax^2 + bx + c$.

The discriminant hypersurface in this example is the swallowtail surface – Fig. 16.23. It has three types of singularities: the swallowtail point, a cuspidal edge and a self-intersection line.

We project the discriminant along the c axis to the plane with coordinates (a, b). The projection of the cuspidal edge to the plane is a semicubic parabola. It is the caustic associated to the reflection group A_3. In the general case one considers any generic projection of the manifold of orbits along any family of curves The final result does not depend on this choice, up to a local diffeomorphism. The *higher-dimensional cuspidal edge* of the discriminant is the variety of the orbits whose points belong to at least two non-orthogonal mirrors.

Caustic of a Group. The *caustic of a reflection group* is the projection of the cuspidal edge of the discriminant along a generic family of lines.

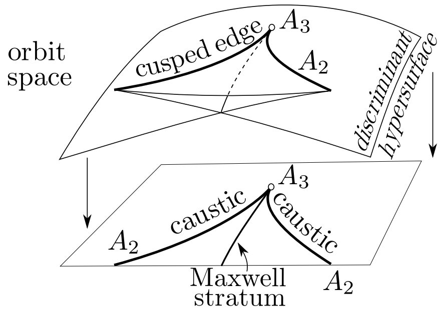


Figure 16.23: The caustic of the reflection group A_3 as the projection of the cuspidal edge of its discriminant.

16.10.4 Caustics and Wave Propagation

In physical terms the caustic is described as the hypersurface swept by the cuspidal edges of moving fronts (Fig. 16.24).

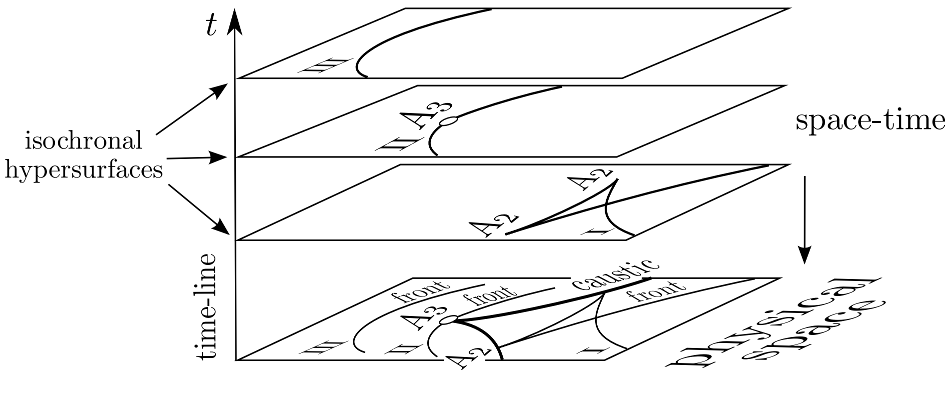


Figure 16.24: The sweeping of a caustic by the cuspidal edges of the fronts.

The propagation of fronts can be described in terms of space-time as a single hypersurface – the union of the momentary fronts belonging to different isochronal hypersurfaces t = const. For the case of a simple singularity corresponding to a reflection group, this hypersurface is locally diffeomorphic to the discriminant of the reflection group.

The cuspidal edges of the momentary fronts sweep the cuspidal edge of the discriminant hypersurface. Its projection into the physical space along the world lines is the surface in physical space swept out by the cuspidal edges of the momentary fronts. Thus the generic lines of the mathematical definition of the caustic of a reflection group are the space-time world lines of the physical description.

Non crystallographic reflection groups. Their discriminants appear in the description of fronts that propagate in a manifold with boundary.

Example. Consider the problem of finding the shortest way of bypassing an obstacle bounded by a smooth hypersurface in Euclidean space $\mathbb{R}^2$ or $\mathbb{R}^3$.

H_3 – If the obstacle is bounded by a smooth curve γ in Euclidean plane, the fronts are the Huygens evolvents (called also involutes) of γ. Near a generic inflection point of the boundary γ, the surface swept out by these moving evolvents in three dimensional space-time is diffeomorphic to the discriminant of the symmetry group H_3 of the icosahedron (Fig. 16.25).

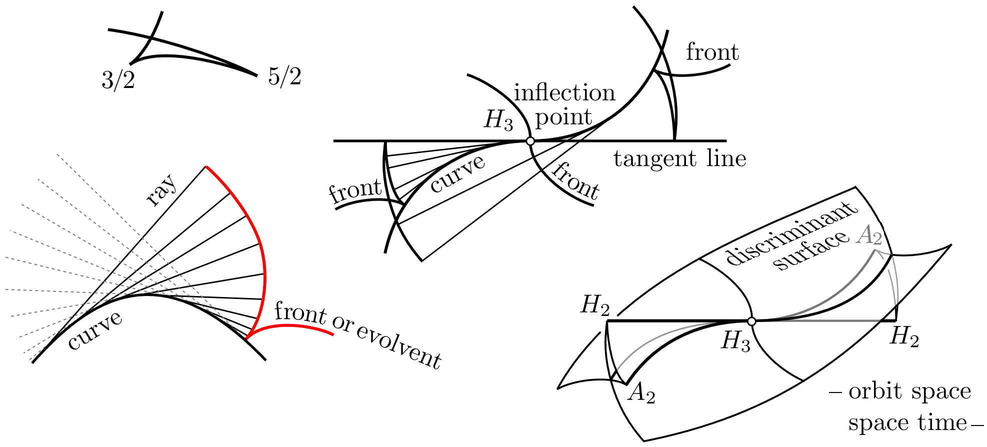


Figure 16.25: Cusps of order 3/2 and 5/2, the evolvents of a plane curve, a perestroika of the evolvents of a curve at an inflection point and the discriminant surface of the icosahedron symmetry group H_3.

H_4 – If the obstacle is bounded by a generic smooth surface in 3-space, each extremal path from a fixed initial point to a variable point of space consists of segments of geodesics on the surface and of segments of tangents to it. The lengths of these extremals define a (multivalued) function of the terminal point. It turns out that the graph of this function has only standard singularities. Near some of them, it is diffeomorphic to the discriminant of the symmetry group H_4 of the hypericosahedron, described on p. 264 (see [112]).

16.10.5 Caustics, Fronts and Stereographic Projections

Poles, polars and Legendre Transform. We start considering a quadric $\mathcal{Q}$ and a point H of a projective space $\mathbb{R}\mathrm{P}^m$. Through this point, move a variable line ℓ that meets the quadric $\mathcal{Q}$ at two points, say, P_1 and P_2. Take the point M on ℓ harmonic conjugate to H with respect to the points P_1, P_2. It is classically known and easy to prove that *the geometric locus of the point M is a hyperplane, say h* (c.f. [110], [107]).

Even when the variable line does not meet $\mathcal{Q}$ (P_1 and P_2 being complex conjugate) M is a real point of the hyperplane h.

Definition. This hyperplane h is called the *polar hyperplane* (or simply the *polar*) of the point H with respect to $\mathcal{Q}$. The *pole* of a hyperplane is the point that admits this hyperplane as its polar (Fig.16.26).

The polar of a point H consists of the poles of all hyperplanes through H.

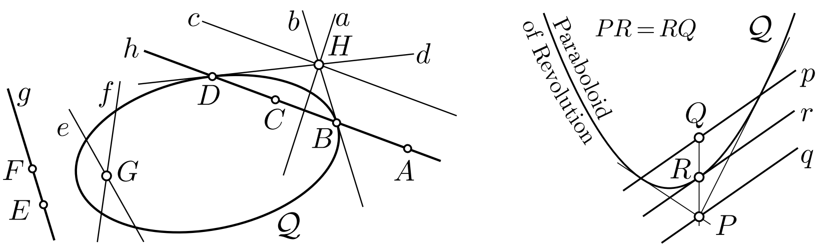


Figure 16.26: The poles $A, B, \ldots$ and their respective polars $a, b, \ldots$.

Polar Duality. This correspondence between hyperplanes of $\mathbb{R}\mathrm{P}^m$ and points of the same space, by means of a quadric, is an isomorphism called *polar duality*

$$\Psi_{\mathcal{Q}} : (\mathbb{R}\mathrm{P}^m)^\vee \xrightarrow{\simeq} \mathbb{R}\mathrm{P}^m .$$

Then the dual of a hypersurface of $\mathbb{R}\mathrm{P}^m$ is represented in the same space!

Definition. The *polar map* Π is the composition of the tangential map with the polar-duality isomorphism Ψ,

$$\Pi = \Psi \circ \mathcal{T} : L_\Gamma \xrightarrow{\mathcal{T}} (\mathbb{R}\mathrm{P}^m)^\vee \xrightarrow{\Psi} \mathbb{R}\mathrm{P}^m .$$

The image of a front $\Gamma \subset \mathbb{R}\mathrm{P}^m$ under the polar map $\Pi = \Psi \circ \mathcal{T} : \Gamma \to \mathbb{R}\mathrm{P}^m$ is called the *polar-dual front* of the initial front i.e. it consists of the poles of all contact elements tangent to Γ.

Affine Quadrics. Given a hyperplane h of $\mathbb{R}\mathrm{P}^m$, the pole of h with respect to $\mathcal{Q}$ is called the *centre* of $\mathcal{Q}$ in the affine space $\mathbb{R}^m = \mathbb{R}\mathrm{P}^m \smallsetminus h$. The hyperplane h is called the "hyperplane at infinity" of $\mathbb{R}^m = \mathbb{R}\mathrm{P}^m \smallsetminus h$.

Hence for a quadric $\mathcal{Q} \subset \mathbb{R}^m$ with centre O *polar duality relates only the hyperplanes not containing O with the points different from O.*

However when $\mathcal{Q}$ is a paraboloid of $\mathbb{R}^m$, its "centre" is in the hyperplane at infinity and, in this case, polar duality is a correspondence between all points of $\mathbb{R}^m$ and all hyperplanes transverse to the axis of the paraboloid.

Exercise. Describe explicitly the polar duality map defined by the paraboloid of revolution $\mathcal{P} = \{(x, \frac{1}{2}\langle x, x\rangle) \in \mathbb{R}^n \times \mathbb{R}\}$, that is, $y = \frac{1}{2}\langle x, x\rangle$.

Answer. The pole of the hyperplane $y = \langle a, x\rangle - b$, with slopes a_i, is the point (a, b). The pole of the hyperplane $y = \langle \widetilde{x}, x\rangle - \widetilde{y}$, with slopes $\widetilde{x}_i$, is the point $(\widetilde{x}, \widetilde{y})$.

Hint. See its geometric construction in Fig. 16.26-right. You can start with $n = 1$.

Proposition. *The polar-dual front of the graph of a function $f : \mathbb{R}^n \to \mathbb{R}$, with respect to the paraboloid $y = \frac{1}{2}\langle x, x\rangle$, is the Legendre transform of f.*

Remark. F. Klein algebraically described Polar duality as follows, [90]. Given two planes $\mathcal{E}$, $\mathcal{E}'$ with respective coordinates (x, y), (x', y'), a linear equation in x, y as well as in x', y'

$$Axx' + B(xy' + yx') + Cyy' + D(x + x') + E(y + y') + F = 0$$

represents a dual transformation between the planes $\mathcal{E}$ and $\mathcal{E}'$. For, if we fix one point in one of the planes, its pair of coordinates being constant, then the equation is linear in the other two coordinates and represents a line in the other plane. The planes $\mathcal{E}$ and $\mathcal{E}'$ play the same role because the equation does not change under permutation of the pairs (x, y) and (x', y'). It makes no difference whether the two planes coincide. *This duality expresses the correspondence pole $\leftrightarrow$ polar with respect to the conic whose equation is*

$$Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0.$$

The tangential map $\tau : \Gamma \to (\mathbb{R}\mathrm{P}^m)^\vee$ of a front $\Gamma \subset \mathbb{R}\mathrm{P}^m$ associates to each point of Γ the hyperplane of the corresponding contact element tangent to Γ (the contact element is a point of L_Γ and the hyperplane is a point of $(\mathbb{R}P^m)^\vee$).

16.11 Symplectic and Contact Topology Topics

16.11.1 Lagrangian intersections - symplectic fixed points

Consider the zero section of the cotangent bundle space of a manifold as a Lagrangian submanifold of this symplectic space. We shall intersect this zero section with a neighbouring Lagrangian submanifold.

Exact Lagrangians. A Lagrangian submanifold of a cotangent bundle space is said to be *exact* if the action form $p\,dq$ on it is exact.

Example. Consider the Lagrangian curve $p = f(q)$ on the surface of the cylinder, which is the cotangent bundle space of the circle $\{q \bmod 2\pi\}$. This submanifold is exact if and only if the mean value of f equals zero.

If the original manifold is simply connected, every neighbouring Lagrangian submanifold is exact.

Suppose that the perturbed exact manifold is a section of the cotangent bundle. Then it is the graph of the differential of a function. The intersection points of the perturbed manifold with the original zero section are then the critical points of this function.

Hence the Morse theory bounds from below the number of intersection points of such perturbed exact Lagrangian submanifolds with the zero section of the cotangent bundle (and hence also of their mutual intersections).

16.11.2 Quasifunctions

Lagrangian intersection theory is a far-reaching generalisation of the Morse bounds to the case of exact Lagrangian submanifolds of the cotangent bundle space which are not sections (see [21], [76], [56], [94]).

Definition. A Legendrian submanifold of the manifold of 1-jets of functions on a compact manifold is a *quasifunction* if it can be connected to the 1-graph of a function by an isotopy in the class of *embedded* Legendrian submanifolds.

Chekanov Theorem. *The projection of a generic quasifunction to the cotangent bundle space intersects the zero section at least b_* times, where b_* is the sum of the Betti numbers of the original manifold.*

Remark. The embeddings in the definition of quasifunctions cannot be replaced by the immersions. This is clear from Fig. 16.27, where both Lagrangian manifolds on the cylinder $T^*\mathbb{S}^1$ are projections of embedded Legendrian curves in the manifold of 1-jets of functions.

The left Lagrangian curve intersects the zero section, while the right one does not intersect it. The left curve can be continuously deformed into the right one in the class of projections of closed Legendrian immersed curves. However, the obvious deformations of this kind contain moments of self-intersection at some intermediate time.

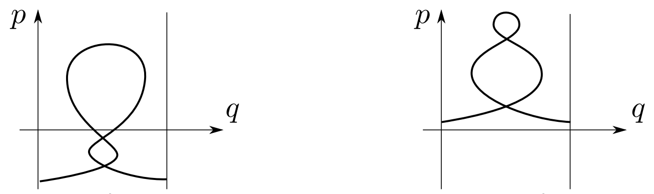


Figure 16.27: A quasifunction and a non quasifunction.

The Chekanov theorem shows that such self-intersections are unavoidable. The left Legendrian curve is a quasifunction, but the right one is not. Any path connecting these curves in the infinite-dimensional space of closed Legendrian immersed curves contains curves with self-intersections.

The Lagrangian intersection theory is closely related to the symplectic fixed-point problem. Indeed, fixed points are points of intersection of the graph of a symplectomorphism with the diagonal of the Cartesian square of the symplectic manifold. Both the graph and the diagonal are Lagrangian submanifolds of the product.

A neighbourhood of the diagonal is symplectomorphic to its cotangent bundle space. The graph is exact (in this neighbourhood) if the symplectomorphism is exact. Hence the symplectic fixed point theorem follows from the Lagrangian intersection theorem, provided that the symplectomorphism does not move any point too far from its original place.

16.11.3 Neutral quadratic forms and their perturbations

All known proofs of the theorems on Lagrangian intersections and symplectic fixed points depend on a generalisation of the Morse theory of geodesics, invented by Rabinowitz (see [108]).

In the classical Morse theory of geodesics the infinite-dimensional variational problem is approximated by a set of finite-dimensional ones. This approximation is based on the fact that the functional one wishes to study is a perturbation of a positive definite quadratic form.

Moreover, this form becomes steeper and steeper when we travel to the higher harmonics in the Fourier representation of the elements of our infinite-dimensional space, while the perturbation remains small in some sense (or at least grows slower than the form at high harmonics).

This makes possible to neglect the higher harmonics completely and to deduce information on the critical points in the infinite-dimensional variational problem from the finite-dimensional Morse theory of the approximated problem.

The new trick invented by Rabinowitz is to apply the same reasoning to quadratic forms which are not positive definite, but which are as far from to be positive definite as possible (forms having as many positive squares in the normal form as negative ones). We shall call them *neutral forms*, leaving this term with no precise definition.

Example. Consider the integral of the action form $p\,dq$ along the maps of the standard circle to the plane with coordinates (p, q).

To parametrise the space of such maps, use Fourier series. Introduce complex notation $z = p + iq$ on the plane, and represent $z(t)$ as the sum of Fourier harmonics $a_k e^{ikt}$. The complex coefficients a_k are the coordinates in our infinite-dimensional space.

The integral of the action form is a quadratic form. A simple calculation shows that it equals the sum of the terms $\frac{1}{2}k|a_k|^2$.

Hence the harmonics with positive "wave number" k contribute with positive squares, and those with negative k contribute with negative squares – as it should be since the integral is the area and positive k corresponds to traversing the oriented boundary of the disc in the positive sense. We see that the quadratic form is steeper in the directions of the higher harmonics, as required. Thus, the non perturbed neutral form has the desired properties.

The fact that the perturbation analysis can be reduced to a finite dimensional problem, neglecting higher harmonics corresponding to large positive and large negative wave numbers, is the infinite-dimensional version of the Hadamard-Perron-Grobman-Hartman-Anosov theorem on the dichotomy of the phase space of a vector field at a stationary saddle point, which is the main fact of the hyperbolic theory of dynamical systems.

The Lyapunov idea of structural stability of the attraction, which is the basis of classical elliptic Morse theory, is replaced in the new theory by structural stability of the neutral saddles of modern dynamical systems theory.

16.11.4 Lagrangian intersections, Floer homology and Casson invariant

The developments of Lagrangian intersections theory led Floer to the eight *Floer homology* groups of the 3-dimensional homology spheres (see [71], [42]).

The idea behind it is that the critical points of a function on a manifold generate the Morse complex, providing the Betti numbers of the manifold.

Lagrangian intersections generate critical points of a functional on an infinite-dimensional manifold. Under certain conditions one can associate to these critical points a generalised Morse complex and its homology. To a homology 3-sphere one can associate certain Lagrangian intersections, and the corresponding homology is an invariant of the homology sphere.

In classical Morse theory, the dimension of the cycle associated to a critical point is the index of the Hessian of the function. In the Floer theory both positive and negative indices are infinite and the dimension is defined only modulo eight.

Floer homology is only defined for 3-dimensional manifolds. One may speculate that there should exist higher-dimensional versions, which could be invariants of contact manifolds of dimension $4n-1$, not simply of smooth manifolds.

Example. The Euler characteristic associated to Floer homology is the Casson invariant of a homology 3-sphere. It counts representations of the fundamental group with appropriate signs.

The *link* of a critical point of a holomorphic function is the intersection of its critical level hypersurface with a small sphere centred at the critical point.

The *Milnor fibre* of a function at a critical point, is the intersection of a nonsingular level hypersurface of the function with a small ball centred at the critical point.

For a function on $\mathbb{C}^3$, the Milnor fibre is a four-dimensional manifold whose 3-dimensional boundary, which is the manifold of our interest, is diffeomorphic to the link of the critical point.

The intersection form on the 2-dimensional homology of the Milnor fibre is symmetric, and its signature is called the *signature of the Milnor fibre.*

Example. For the function $x^2+y^3+z^5$, the link is the Poincaré dodecahedral space, which is a homology 3-sphere. Its Milnor fibre is a 4-manifold bounded by a dodecahedral space. The intersection form on its two-dimensional homology is negative definite and the signature is equal to -8.

This form is, with a minus sign, the restriction of the Euclidean scalar product to the lattice generated by the vectors with scalar square 2, whose angles are defined by the Dynkin diagram E_8.

Wahl and Neumann (see [105]) have discovered an astonishing relation between the topology of critical points of holomorphic functions and the theory of Floer homology.

Theorem 11. *For any weighted-homogeneous function of three complex variables, whose link is a homology* 3*-sphere, the Casson invariant (up to a universal factor) equals the signature of the Milnor fibre.*

Remark. The link of the E_8 singularity in five variables

$$x_1^2 + x_2^2 + x_3^2 + y^3 + z^5 = 0$$

is homeomorphic but not diffeomorphic to the 7-sphere. This smooth manifold is one of the 28 Milnor *exotic spheres*. The above simple equation for that Milnor sphere was discovered by Brieskorn.

Unlike the Casson invariant, the signature is defined for any singularity. One may ask whether the Casson invariant and Floer homology can be defined in this situation.

The link has a natural contact structure and the Milnor fibre a symplectic structure. Its homology is generated by its vanishing cycles, which are Lagrangian spheres. The intersection form is related to the linking of their Legendrian representatives in the link manifold.

The Morse complex corresponding to this situation, has yet to be constructed for more general singularities than those considered by Wahl and Neumann.

16.11.5 Characteristic classes in quantisation conditions

The existence of symplectic topology was first appreciated by the mathematical community when the characteristic class entering into the quantisation conditions was discovered. V. Arnold christened it as the *Maslov class* in [5].

In its simplest version this class is a one-dimensional cohomology class of a Lagrangian submanifold of the cotangent bundle space of a manifold.

The case where the manifold is $\mathbb{R}^n$ and the Lagrangian submanifold lives in the standard symplectic phase vector space $\mathbb{R}^{2n}$ is very instructive, even for $n = 1$ (curves in the phase plane).

The Maslov class associates to any closed curve of a Lagrangian submanifold an integer, its *Maslov index*, which depends only on the homology class of the curve.

The Maslov index of a curve is the intersection index of the curve with the critical set of the projection of the Lagrangian submanifold to the base space of the fibration, that is, it is the intersection index of the curve with the preimage of the caustic. It is well defined because the critical set is a cycle of codimension one in the Lagrangian manifold and has a natural co-orientation. The definition of this co-orientation was given in [5].

Problem. Guess the general definition of the above natural co-orientation from the co-orientations depicted by arrows in the example of Fig. 16.28.

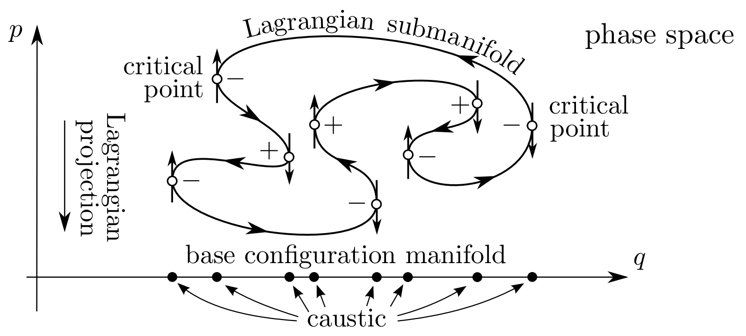


Figure 16.28: The natural co-orientation of the critical set and the Maslov index.

In physical terms the Maslov index describes the well-known effect of the loss of a quarter of a wave at a caustic. This correction to the short wave approximation produces the 1/2 correction term in the Bohr-Sommerfeld quantisation conditions.

These conditions describe the possible energy levels in a one-dimensional quantum system in terms of the area bounded by the corresponding phase curve on the phase plane, that is, in terms of the integral of the action form.

In some units, called *quanta of action* and depending on the Plank constant, the area should be an integer if the correction is not taken into account.

The correction due to the existence of the caustic is 1/4 for each crossing of a critical point of the projection in the direction defined by the natural co-orientation. Hence the total correction is 1/4 multiplied by the Maslov index of the phase curve.

Since this index equals two, according to Fig. 16.28, the number of quanta of action inside the n-th energy level should be $n + 1/2$. This is the quantisation condition.

It was known to physicists that something similar happens in higher dimensions, at least in some classical integrable systems. Maslov has tried to formulate it mathematically and has recognised the topological nature of the correction term as well as the importance of Lagrangian submanifolds in this context.

By the way, in the first Arnold work on symplectic topology ([3]) he called them *null-submanifolds*. But Maslov, disliking that name, advised Arnold to rename them after a mathematician he likes, and christened these submanifolds *Lagrangian* because the so-called Lagrangian brackets (the 18th-century way to describe the symplectic structure) vanish there. Then, following Maslov advise, Arnold christened Legendrian submanifolds after Legendre and the Maslov class after his friend.

Using the short-wave approximation, Maslov has proved in his thesis [101] that his co-orientation rule defines a cohomology class modulo four. Reference [5], which explains why it is an integer-valued class, was initially written as a referee report.

Remark. The short wave or quasiclassical approximation is now usually called the WKB(J)-method in physics, by the names of the quantum mechanics people using it. It seems that the method was first published by Carlini (1817) [55]. A detailed exposition of the method (later used by Stokes, Kelvin, and many others in the 19th century) was published by Jacobi (see Vol. 7 of his Collected Works, pp. 175-245 – we are indebted to Professor S. Graffi for these references).

The Maslov index of a closed curve on a Lagrangian submanifold in $\mathbb{R}^{2n}$ can be defined as the number of complete rotations, in the unit circle of the complex numbers, of the square of the determinant of a unitary matrix described below.

Consider the complex space $\mathbb{C}^n$ with its usual Hermitian structure. The real part of the Hermitian structure is a Euclidean structure in $\mathbb{R}^{2n}$, its imaginary part is a symplectic structure, and these two structures are invariant under multiplication by i. The Lagrangian subspaces are the real n-subspaces orthogonal to their images under multiplication by i.

Each orthonormal frame in a Lagrangian subspace defines a Hermitian orthonormal frame in the complex space. Consider the unitary operator sending the fixed standard coordinate unitary frame to the chosen orthonormal frame in a Lagrangian subspace, and take the determinant of its matrix.

Theorem 12. *The square of this determinant is independent of the choice of the frame in the given Lagrangian subspace.*

Proof. Any change of this frame multiplies the unitary operator by an orthogonal real one. The determinant of a real orthogonal operator is $+1$ or -1. Hence, the square of the determinant remains unchanged when the frame changes. □

The tangent spaces of a Lagrangian submanifold of Euclidean phase space are Lagrangian subspaces of this phase space.

Theorem 13. *The Maslov index of a curve on a Lagrangian submanifold of Euclidean phase space is equal to the number of rotations of the square of the determinant corresponding to the tangent spaces of the Lagrangian submanifold along the curve.*

16.11.6 Lagrangian and Legendrian characteristic classes

The Maslov class, dual to the caustic of a Lagrangian map, can be extended to a more general category of real vector bundles whose complexifications are trivial (see [3]). There exist other natural generalisations of the preceding construction too.

Definition. The *Lagrangian Grassmannian* is the manifold of the Lagrangian subspaces of a symplectic real vector-space.

As we have seen in Section 3.15, the Lagrangian Grassmannian for the symplectic real vector-space $\mathbb{R}^{2n}$ is the homogeneous space $\Lambda(n) = \mathrm{U}(n)/\mathrm{O}(n)$. Its cohomology classes induce cohomology classes on the Lagrangian submanifolds of $\mathbb{R}^{2n}$, called *Lagrangian characteristic classes*.

Indeed, any Lagrangian submanifold of the vector-space is sent to the Lagrangian Grassmannian by the *Lagrangian Gauss map* which associates to a point of a Lagrangian submanifold the direction of its tangent space. The pull-back of a class of the Grassmannian is a characteristic class of the Lagrangian submanifold.

Example. The Maslov class is the pull-back of the *universal Maslov class* – the one-dimensional cohomology class of the Lagrangian Grassmannian, induced from the unit circle by the map $\det^2$.

The cohomology ring of the Lagrangian Grassmannian is well-known [53]. For the purposes of this book it suffice to know that such cohomology ring exist, is known and is used to get invariants of Lagrangian cobordism classes, but we will not describe explicitly its structure. We only mention that the multiplicative generators in dimension $4k+1$ are related to the Pontryagin classes by the Bott periodicity shift.

The one-dimensional class is dual to the set of critical points of a Lagrangian map. This set is generically stratified according to the classes of singularities of caustics of different codimensions: cusps or cuspidal edges, swallowtails, etc.

One can construct higher-dimensional characteristic classes dual to higher codimension Lagrangian or Legendrian singularities. These constructions provide a lot of geometric information on the coexistence and interrelations of singularities of caustics and wave fronts.

Example. The number of swallowtails on a generic caustic of a compact Lagrangian 3-manifold or on a generic front of a compact two-dimensional Legendrian manifold is even. Indeed, a swallowtail is an endpoint of the self-intersection line of a caustic or of a front, while at any other singular point on the self-intersection line there meet an even number of rays.

We refer to [128] for dozens of less trivial examples. Since in dimensions higher than six there is still no explicit description of the natural stratification of the space of Lagrangian or Legendrian singularities, the combinatorics related to the Lagrangian and Legendrian singularities defining higher characteristic classes is an immense challenge.

The big progress in the beginning of the 90's, due to Vassiliev and Kontsevich, on a similar problem in knot theory, shows that the complicated combinatorics may hide rather simple and universal algebraic structures. Last years, this theory, known as Vassiliev invariants theory, has had an extremely wide development with thousands of publications. In chapter 18, we describe the initial simple and clear ideas of Vassiliev's theory. In fact this progress in knot theory was initially a byproduct of previous work on global theory of Lagrangian and Legendrian singularities.

16.11.7 Lagrangian and Legendrian cobordisms

Consider the cotangent bundle of a base manifold with boundary.

The *Lagrangian boundary* of an immersed Lagrangian submanifold is a Lagrangian submanifold immersed in the cotangent bundle of the boundary of the base manifold: by definition, it is the projection of the intersection of the Lagrangian manifold with the boundary of the phase space along the characteristics of this boundary.

Physically an immersed Lagrangian submanifold represents the short-wave approximation to a wave state inside the base manifold. If the base manifold has a boundary, the wave state inside the base manifold determines the state on the boundary (think of the light inside a room and on its walls). Hence, it is natural that the interior Lagrangian manifold defines a *Lagrangian boundary*, which is an immersed Lagrangian submanifold of one lower dimension.

One may also imagine the restriction of a (multivalued) function to the boundary of its domain.

Definition. Two immersed Lagrangian submanifolds L_0, L_1 of the cotangent bundle space T^*B are (cylindrically) *Lagrangian cobordant*, if the difference between $L_1 \times 1$ and $L_0 \times 0$ is the Lagrangian boundary of an immersed Lagrangian submanifold L of $T^*(B \times [0,1])$ which intersects transversely the boundary of this cotangent bundle (see Fig. 16.29, where π is the projection of the hypersurface $\partial\big(T^*(B \times [0,1])\big)$ along its characteristics). We call L a *Lagrangian cobordism* between L_0 and L_1. In the oriented version, as usual, $\partial L = L_1 - L_0$, where the minus means the reversal of orientation.

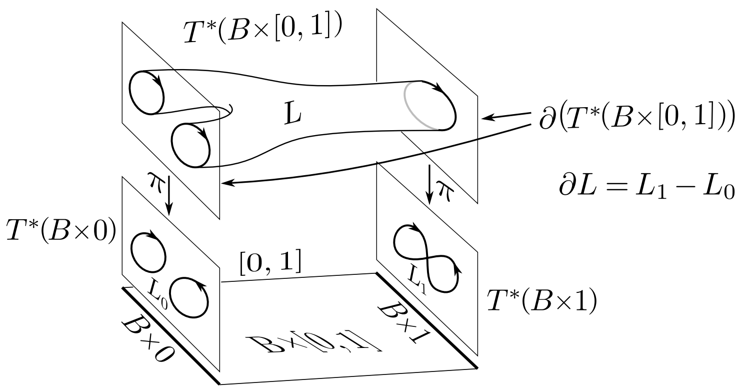


Figure 16.29: An (oriented cylindrical) Lagrangian cobordism.

Example. Two oriented closed curves immersed in the plane are Lagrange (cylindrically, orientable) cobordant if and only if they have the same Maslov index and bound equal areas (have equal action integrals).

As usual, one defines addition as the disjoint union, forming a (commutative) semigroup from cobordism classes. This semigroup is in fact a group.

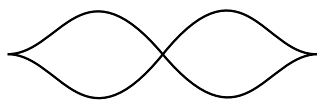
Example. The group of Lagrangian (oriented cylindrical) cobordism classes of plane curves is $\mathbb{Z} \oplus \mathbb{R}$.

The characteristic numbers of cobordant objects are equal. So, one can use the numbers of singular points of Lagrangian maps to distinguish cobordism classes or use information on the classes to understand adjacencies of singularities.

There exists a similar theory in contact geometry for Legendrian singularities and cobordisms. In this case one may consider front cobordisms, defined as cobordisms of stratified varieties – they faithfully represent the Legendrian cobordisms of immersed Legendrian subvarieties in contact spaces.

Theorem 14. *The Legendrian (oriented cylindrical) cobordism group of Legendrian curves immersed in the standard contact* 3*-space is isomorphic to the group of integers* $\mathbb{Z}$.

The generator is the class of the Legendrian curve whose front is the bow-tie curve, shown in Fig. 16.30.



$q = \sin t, \quad p = 8\cos 3t,$
$z = \sin 4t + 2\sin 2t$

Figure 16.30: The bow-tie.

We leave to the reader the pleasure of finding the geometric proof of this fact: one can really decompose any front having no vertical tangents into bow-ties. Our fronts have no vertical tangents, since our 3-space with coordinates $(p, q; z)$ carries the standard contact structure $dz = p\,dq$.

The front lives on the plane with coordinates z and q. It is non vertical since its inclination p is finite.

The number of bow-ties in the decomposition can be read immediately from the given generic front. Recall that a generic front has only cusp singularities and that our fronts are co-oriented (since they have no vertical tangents). Choose the co-orienting one-form $dz = p\,dq$.

Definition. A cusp of an oriented and co-oriented front is *positive* (*negative*) if the orienting motion leaves the cusp point in the direction to the side where the co-orienting one-form is positive (negative).

Example. Both cusp points of the bow-tie front of Fig. 16.30 are positive. For the bow-tie with the opposite orientation, but same co-orientation, the cusps are negative.

Theorem 15. *The difference between the numbers of positive and negative cusps on an oriented and co-oriented front is an invariant of the oriented Legendrian cobordisms.*

Hence the number of the standard bow-ties to which the front is cobordant, is easily computable. It is equal to one half of the above difference.

Remark. The difference between the numbers of positive and negative cusps is called *the Maslov index* of the front (or of the Legendrian curve). It is indeed the Maslov index of a curve on a Lagrangian submanifold of the space of a Lagrangian fibration. To obtain these objects one should symplectise the given Legendrian and contact objects.

The Maslov index is the only invariant of the oriented Legendre cobordism of oriented and co-oriented fronts in the plane.

Theorem 16. *The non oriented Legendrian curves immersed in the standard* 3*-space form a trivial cobordism group.*

Indeed, the bow-tie cobordism to the void curve is shown in Fig. 16.31. The Legendrian surface represented by this sequence of transformations of the front, is a Möbius band, bounding a circle (whose projection is the bow-tie). Any front is cobordant to zero because it is cobordant to several bow-ties.

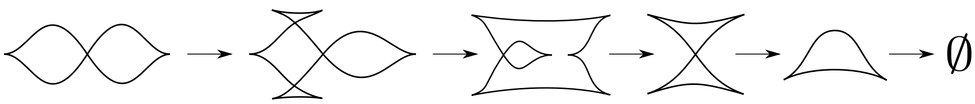


Figure 16.31: The non-oriented Legendrian cobordism of the bow-tie to zero.

The Legendrian cobordism groups are simpler than the Lagrangian ones, since the invariants are discrete, unlike the area in the Lagrangian case. We quote without proof from the book of Audin [43]:

Theorem 17. *The Legendrian oriented cobordism groups of the standard Euclidean spaces form a graded anti-commutative ring. This ring (tensored with the rationals to kill torsion) has one free multiplicative generator in degree $4k+1$ for each $k = 0, 1, 2, \dots$.*

The non-oriented Legendrian cobordism groups form the ring

$$\mathbb{Z}[x_5, x_9, x_{11}, \dots]$$

(one generator of each odd degree except those of the form $2^k - 1$).

The Legendrian cobordism rings formed by Legendrian submanifolds of the manifold of co-oriented contact elements $ST^*\mathbb{R}^m$ are naturally isomorphic to the rings of the theorem above (describing the Legendrian submanifolds of the spaces of 1-jets of functions on Euclidean spaces).

16.11.8 Lagrange embeddings and inclusions

A circle can be embedded into the plane and hence there exists an embedded Lagrangian torus in the standard phase space $\mathbb{R}^4$. However, this embedding is not exact because the integral of the action form is a multivalued function on the torus. This is a particular case of the following general fact:

Gromov Theorem. *No compact smooth manifold admits an exact Lagrangian embedding into the standard symplectic space $\mathbb{R}^{2n}$.*

This theorem has a simple topological proof for all surfaces different from the torus, for which the theorem is really very hard and was proved only recently by methods from the theory of (pseudo) holomorphic functions.

This technique was introduced into symplectic topology by Gromov ([76]). The most striking results in symplectic and contact topology have been so far obtained only by this method, which seems foreign to the subject. Its strange power is a cousin of the use of variational methods in elliptic PDE's.

Definition. A *Lagrangian inclusion* is a smooth map of a manifold into a symplectic manifold, which is a Lagrangian embedding in a neighbourhood of almost every point (and which therefore induces the zero 2-form from the symplectic structure).

Example. The *conormal bundle* of a subvariety in configuration space is the variety of cotangent vectors of the configuration space which vanish on the tangent vectors of the subvariety. The *Givental open umbrella singularity* is the singularity of the conormal bundle space of a semicubic parabola in the plane. The Givental singularity is the image of a Lagrangian inclusion (of the plane). Topologically, this surface of $\mathbb{R}^4$ is non-singular. But it has one point of non smoothness.

Indeed, the Givental surface is parametrised by two parameters: t along the semi-cubic parabola and s across it, namely

$$q_1 = t^2 \ , \quad q_2 = t^3 \ , \quad p_1 = -3st \ , \quad p_2 = 2s \ .$$

Theorem 18 (Givental, Ishikawa, Zakalyukin)**.** *The only singularities of the generic Lagrangian inclusions of a surface in space are transverse self-intersections and Givental singularities. These points are stable.*

Theorem 19 (Givental, [79])**.** *Any orientable surface of genus $g > 1$ admits a Lagrangian inclusion into the standard symplectic Euclidean 4-space with $2g - 2$ Givental open umbrella singularities and no self-intersections* (topologically this inclusion is an embedding)*.*

Any non orientable surface of negative genus $-4k$ admits a Lagrangian embedding into the standard symplectic Euclidean 4-space with no singularities at all.

The exact Lagrangian inclusion of the real projective plane into $\mathbb{R}^4$ is represented in Fig. 16.32 by its front.

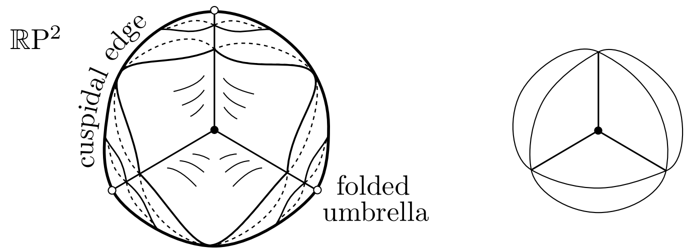


Figure 16.32: An exact Lagrangian inclusion of the projective plane into 4-space.

This front is a surface in $\mathbb{R}^3$ which has a semi-cubic cuspidal edge and a self-intersection line, like fronts of smooth Legendrian manifolds have. The cuspidal edge and the self-intersection line meet at three points. The singularities of the front at these points are called *folded umbrellas*. At a neighbourhood of a folded umbrella the front can be described by the normal form $y^2 = z^3x^2$ – Fig. 16.33.

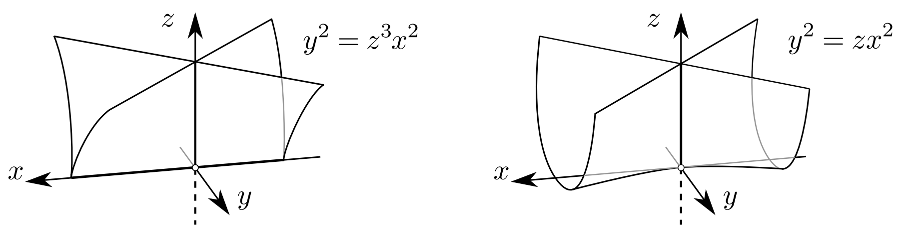


Figure 16.33: The folded umbrella and the Whitney-Cayley umbrella.

Like the ordinary Cayley-Whitney umbrella $y^2 = zx^2$, the folded umbrella contains a handle $(x = y = 0, z < 0)$, shown in Fig. 16.33 as a dotted line. The handle is not included in the real front.

As we have seen in Chapter 15 (pp. 553-554), the Cayley-Whitney umbrellas are the only singularities of generic maps of surfaces into 3-space, besides transverse crossings of two or even three branches. In a neighbourhood of such a singular point, the map can be written in the Whitney normal form $(x = u, y = uv, z = v^2)$, where u and v are local coordinates on the surface.

The folded umbrella is a standard element of many singularities in symplectic and contact geometries.

Example. The tangent lines of a generic smooth curve in the Euclidean 3-space sweep out a surface. The original curve is a semi-cubic cuspidal edge of this surface. At points where the torsion of the curve vanishes, the surface is locally diffeomorphic to a folded umbrella.

Folded umbrellas and double points obstruct the smooth Lagrangian embedding of surfaces into Euclidean space. One cannot eliminate them by deforming the embedding. However, two folded umbrellas may replace one double point (see Fig. 16.34).

Each part of Fig. 16.34 represents a front in 3-space. Each front consists of two surfaces $z = f(x, y)$, $z = -f(x, y)$ connected along the cuspidal edges, where $z = 0$. The shape of the graph of f is shown by the level lines and by the directions of fastest descent.

The Lagrangian submanifold corresponding to the left front has no self-intersections. Indeed, the planes tangent to the front at any two points with equal coordinates (x, y) are nowhere parallel because they are symmetric with respect to the plane $z = 0$ and are not horizontal, since the function has no critical point outside the cuspidal edge.

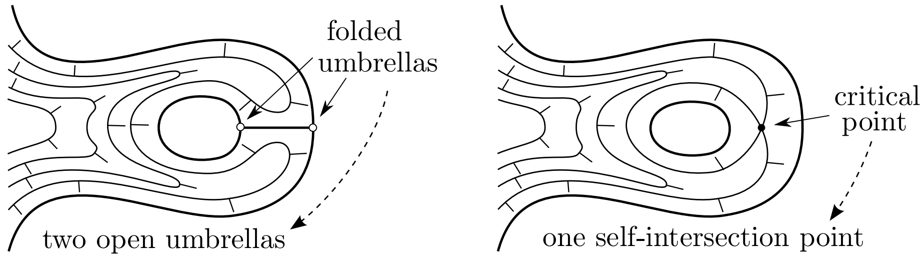


Figure 16.34: The Givental handles.

This cuspidal edge contains two folded umbrellas and the corresponding Lagrangian submanifold of $\mathbb{R}^4$ has two open umbrellas.

The Lagrangian submanifold associated to the right front has one self-intersection point, which corresponds to the critical point of f. It has no Givental open umbrellas.

16.11.9 Lagrangian and Legendrian knots

A smooth manifold may have many different embeddings into a Euclidean space. Two smooth embeddings are called *isotopic* if they belong to the same connected component of the space of embeddings. The classification of these components is a generalisation of knot theory, which studies the embeddings of the circle into 3-space.

An equivalent (and in a sense more "real world") definition of a knot starts from a standard straight line in 3-space. A knot is represented by an embedding which differs from the standard one only in a finite ball (Fig. 16.35).

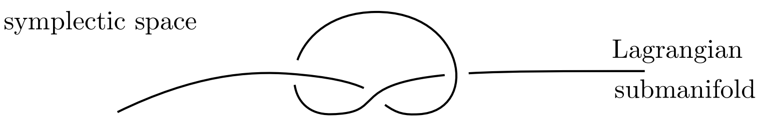


Figure 16.35: A knot on an infinite string.

Definition. A *Lagrangian knot type* is a connected component of the space of Lagrangian embeddings of $\mathbb{R}^n$ into the standard symplectic space $\mathbb{R}^{2n}$, differing from the embedding of a fixed Lagrangian plane only inside finite balls. The *trivial type* is the one containing the standard Lagrangian n-plane.

Problem ([28]). Do there exist nontrivial Lagrangian knot types?

Theorem 20 ([67])**.** *All Lagrangian knots in the standard symplectic* 4*-space are trivial in the sense of differential topology.*

In other words, any topological Lagrangian 2-plane which is "standard" outside some ball is unknotted: one can connect the plane with the perturbed surface by a continuous deformation in the class of smooth (not necessarily Lagrangian) embeddings.

Moreover, this deformation can even be realised by a time-dependent flow, fixing the points outside some ball and transforming the plane into the perturbed surface at time 1. If this flow were Hamiltonian, all Lagrangian knots would be trivial. The theorem above means that nontrivial Lagrangian knots, if they do exist, are purely Lagrangian – topologically they are trivial.

Whether higher-dimensional Lagrangian knots can be topologically nontrivial is unknown.

A Legendrian curve in the standard contact $\mathbb{R}^3$ or $\mathbb{S}^3$ may be considered as a knot in the usual sense, and every knot has Legendrian representatives. However the type of the knot in the usual sense does not define its Legendrian type (the component of the space of Legendrian embeddings).

Example. The Legendre curve whose front is the bow-tie is in the same trivial class of ordinary knots as the "lips" curve – Fig. 16.36. But they are different as Legendrian knots in $\mathbb{R}^3$. Indeed, the "lips" curve is Legendrian cobordant to zero, while the bow-tie is not.

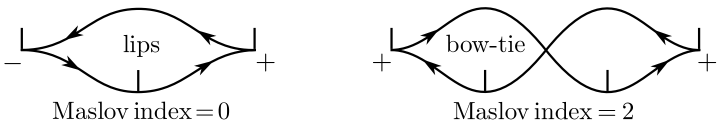


Figure 16.36: The fronts of two different purely Legendrian knots that are unknotted in the usual sense.

The Maslov index, distinguishing Legendrian cobordism classes, is not the only invariant of Legendrian isotopies of curves, even if they are unknotted in the ordinary sense.

Definition. The *Bennequin invariant* of a Legendrian curve in $\mathbb{R}^3$ is the intersection index of an oriented surface bounded by the curve with a curve obtained from the given one by a small shift in the direction orthogonal to the contact plane.

Example. The Bennequin invariant of the curve shown in Fig. 16.37 is $b = 1 + s + 2t = 11$.

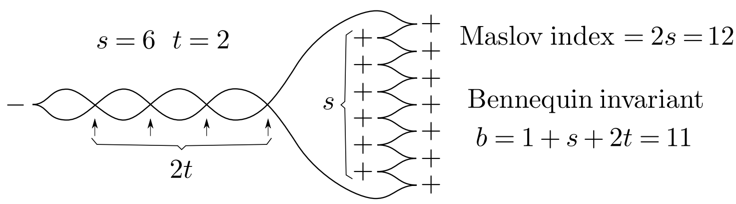


Figure 16.37: The classification of the purely Legendrian knots in $\mathbb{R}^3$.

Bennequin Theorem ([48])**.** *The value of the Bennequin invariant on an unknotted curve in the standard contact* $\mathbb{R}^3$ *is positive.*

Remark. The Bennequin invariant may be defined for higher-dimensional Legendrian submanifolds of topologically trivial contact manifolds. In a sense, it is a quadratic form. For instance, this self-linking number is multiplied by 4 if the curve is covered twice. The Bennequin inequality claims that this form is positive definite.

Bennequin used this inequality to prove that a certain twisted contact structure in $\mathbb{R}^3$ is exotic. Namely, he constructed for the twisted contact structure a Legendrian curve, unknotted in the ordinary sense, whose Bennequin invariant (depending on the contact structure) is not positive.

Eliashberg Theorem ([66])**.** *The Maslov index and the Bennequin invariant are the only invariants of Legendrian isotopies of topologically unknotted Legendrian curves in the standard contact space* $\mathbb{R}^3$.

The fronts of the Legendrian curves representing all the classes are similar to that shown in Fig. 16.37: To obtain all classes, one has just to consider both orientations of the fronts with all possible s, $t = 0, 1, \ldots$.

16.11.10 Classification of Contact Structures

The first example of an exotic contact structure in $\mathbb{R}^3$ was constructed by Bennequin. Now all such structures are known.

Theorem 21 (Eliashberg, [68])**.** *The complete list of non-equivalent contact structures on* $\mathbb{R}^3$ *is countable.*

The list itself can be found in [68]. The theorem should be compared with the fact, due to the same author, that there exists a *continuous family* of pairwise non-equivalent contact structures on the filled torus $\mathbb{S}^1 \times \mathbb{R}^2$.

The contact structures on a closed manifold are rigid: Any small deformation is equivalent to the initial contact structure by a diffeomorphism close to the identity (J. Gray, [75]).

The homotopy type of the field of hyperplanes defining a contact structure is of course an invariant of the structure.

Any closed 3-manifold has a contact structure (Martinet, [98]).

The fields of tangent hyperplanes on $\mathbb{S}^3$ are classified by the integers. As we have seen, this sphere is the group $\mathrm{Spin}(3) \approx \mathrm{SU}(2)$. Thus, we identify all the tangent spaces using left translations on the group. A field of hyperplanes is then described by a map $\mathbb{S}^3 \to \mathbb{S}^2$. The homotopy classes of the fields of hyperplanes are labelled by the Hopf invariants of these maps.

Theorem 22 (Eliashberg, [65])**.** *Every class of fields of planes on $\mathbb{S}^3$ contains (up to isotopy) exactly one contact structure, with the exception of the zero class, which contains exactly two contact structures.*

Ginzburg ([77]) defined some natural contact cobordism groups. He proved that all contact cobordism groups of contact manifolds of dimensions $4k + 3$ are trivial, while the contact cobordism group of contact manifolds of dimension $4k + 1$ consists of two elements.

Example. A contact circle* is not contact cobordant to zero. This simply means that a compact oriented surface bounded by a circle has no line fields transverse to the boundary circle. The union of two circles is contact cobordant to zero, however, since there exists a field of lines on the cylinder transverse to both boundary circles.

In the *symplectic cobordism* theory of Ginzburg, the situation is quite different. The invariants of symplectic cobordism are the integrals of certain differential forms, which are products of the Chern forms and of the exterior powers of the symplectic structure. Hence the cobordism groups are not finitely generated.

Example. The only cobordism invariants of symplectic structures on surfaces are the symplectic area and the Euler characteristic, which of course are additive.

*That is, a circle provided with the trivial contact structure

16.11.11 Existence of symplectic and contact topologies

Most natural questions in symplectic and contact topologies are yet unanswered, but one thing is firmly established: symplectic and contact topologies *do* exist. This statement is not a noticing of mathematical activity in these domains, but a technical result: A highly nontrivial theorem.

The definitions of symplectic and contact structures use the differentiable structure of the underlying manifold. Topological symplectic or contact geometry should be invariant under homeomorphisms preserving the structure.

Example. A closed curve dividing the 2-sphere into two parts of equal areas intersects any equator at least at two points. This is true for non smooth Jordan curves. Hence the fact belongs to topological symplectic geometry.

Similar examples show that topological volume-preserving geometry does exist. However, the symplectic theory is more delicate when the dimension is higher than 2.

Instead of pathological topological objects one may consider limits in the C^0 topology (where no convergence of the derivatives is supposed) of smooth objects, like Lagrangian submanifolds, symplectomorphisms, etc.. The problem arises whether these limits preserve the properties of the smooth original objects.

Of course, the limit should not be *too* bad. For instance, the limit of a sequence of Lagrangian curves might be the whole phase plane, which of course should not be considered as a honest topological Lagrangian submanifold.

The resulting topological theory would preserve the traces of the symplectic (and not just volume-preserving) geometry only if the topological limit does not change their symplectic status when they are smooth submanifolds or maps.

Theorem 23 ([76])**.** *A diffeomorphism of a compact symplectic manifold is a symplectomorphism, if it is the C^0 limit of symplectomorphisms.*

Theorem 24 (Laudenbach-Sikorav, 1993)**.** *A smooth embedding of a closed manifold into the standard symplectic space is Lagrangian if it is the C^0 limit of Lagrangian embeddings of the same closed manifold.*

Similar results hold for contactomorphisms and Legendrian submanifolds.

In spite of the recent proof of these long awaited theorems, the construction of topological symplectic and contact geometries (even of the PL-version, which is important in optimal control theory) is yet to be achieved.

Remark. We call *topological symplectic and contact geometries* the studies of symplectic and contact properties of non smooth objects, calling *symplectic and contact topologies* the studies of discrete invariants of smooth objects in symplectic and contact manifolds (see our discussion in Sect. 15.10).

16.11.12 Contact and symplectic worlds

Symplectic and contact geometries are of course differential geometries of manifolds with some additional structures. Some rather natural axioms led Cartan to a small list of natural geometries of this kind, associated with the simple (pseudo)-groups of diffeomorphisms.

The Cartan list of simple pseudo-groups contains real and complex differential and volume-preserving geometries, symplectic and contact geometries, and a few conformal versions of the preceding geometries. This list is somewhat similar, and closely related, to the Killing list of simple Lie algebras mentioned above and whose Dynkin diagrams were discussed and described in Chapter 7, pp.260-265 (see also Fig. 16.9).

It is well known in the theory of Lie algebras that practically any fact of matrix theory can be reformulated in coordinate-free terms of the so-called root theory, which is a natural extension of the theory of the eigenvalues. The roots of the linear group form the simplest Dynkin diagram A_k.

Once the result is formulated in terms of the roots, it becomes also meaningful for the other Dynkin diagrams. In this way one obtains at least a conjecture valid for all simple Lie algebras. In most cases one can prove the conjecture, though sometimes slightly correcting it.

This method unifies the geometries of finite-dimensional simple Lie algebras. It seems that a similar unification may be useful in the infinite-dimensional case of the simple groups of diffeomorphisms.

From this point of view, symplectic, contact or holomorphic geometries and topologies should be considered as *sisters* of ordinary geometry and topology rather than as parts of them, in the same way as in the theory of simple Lie groups one considers the orthogonal group as a brother of the linear group rather than as its subgroup (in spite of the fact that the orthogonal group *is* a group of linear transformations preserving an additional structure). One may thus imagine that quite a few of the notions of ordinary geometry and topology of manifolds may have parallels in symplectic, contact, complex, and other geometries.

Example. The following list of parallel objects in the real and in the complex geometries is well known (see e.g. [21]) :

$\mathbb{R}$	$\mathbb{C}$
$\mathrm{O}(n)$, $\mathrm{SO}(n)$	$\mathrm{U}(n)$, $\mathrm{SU}(n)$
π_0	π_1
$K(\pi, 0)$	$K(\pi, 1)$
$\mathbb{Z}_2$	$\mathbb{Z}$
$\mathbb{R}\mathrm{P}^n$	$\mathbb{C}\mathrm{P}^n$
Stiefel-Whitney classes	Chern classes
$\mathbb{S}^1 = \mathbb{R}\mathrm{P}^1$	$\mathbb{S}^2$
$\mathbb{S}^1 = \mathbb{R}/\mathbb{Z}$	$\mathbb{S}^3$
Morse theory	Picard–Lefschetz theory
symmetric group	braid group
boundary	ramified covering
orientation of $\mathbb{R}^n$	element of $\pi_1(\mathrm{U}(n))$.

It is clear that the complexification of "homology" is not at all homology theory in the sense of homological algebra, but some algebraically completely different object, yet to be discovered.

Similarly to the complexification, the symplectisations and contactisations of the usual geometric objects may be very different from the original objects. There exist no axioms yet for mathematical operations of this rank, like quantisation, superisation, symplectisation.

However one may guess that in many cases symplectisation can be achieved by the following procedure. One starts from some object in an ordinary manifold. One symplectises it, considering the cotangent bundle space and associating to the original object some kind of prolongation (the cotangent bundle space of a submanifold, the action of a diffeomorphism on cotangent vectors, etc.). Finally one can try to generalise the properties of the resulting objects to make them symplectically invariant.

In such a way one obtains, in the preceding examples, Lagrangian manifolds or varieties starting from ordinary manifolds or varieties, Hamiltonian vector fields from ordinary vector fields, and so on.

It seems that the Morse number and the Lagrangian intersection theory have something to do with the symplectisations of the Euler characteristic and of the Lefschetz fixed point theorem of ordinary topology, while Lagrangian and Legendrian cobordism theories might be the symplectisation and the contactisation of the usual cobordisms.

Most of the branches of mathematics (from the theory of PDEs to the calculus of variations, from the theory of group representations to number theory) have been symplectised or are under symplectisation currently.

In this brave new symplectic world the old concrete theories live a new life in the company of cousins from whom they were separated before symplectisation, similarly to the ellipse, the parabola, and the hyperbola, which were separated before projectivisation.

One starts from some concrete theory. Say, one considers the elementary theory of the Legendre transformation or the geometry of the equidistant surfaces of a surface in Euclidean 3-space, or the classical theory of pedal surfaces (also called derivative surfaces), see [72].

The *derivative* of a hypersurface in Euclidean space is the variety formed by the feet of the perpendicular from the origin to the tangent plane of the given hypersurface – Fig. 16.38. The *primitive* hypersurface of a given hypersurface is the envelope of the hyperplanes orthogonal to the radius vectors of the points of a given hypersurface (see [81], Fig. 16.38).

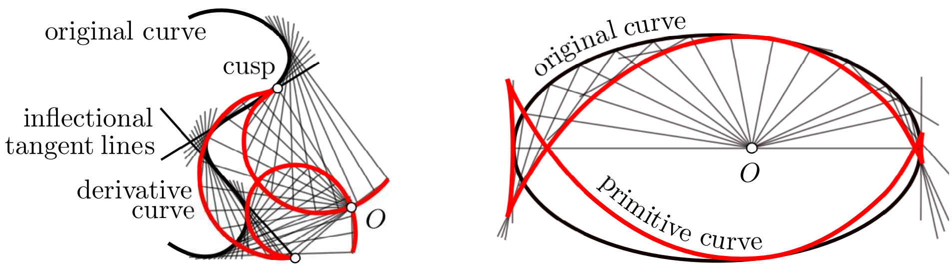


Figure 16.38: The derivative and the primitive of a hypersurface.

In contact geometry these hypersurfaces show their true face: They are fronts of Legendrian maps. It follows that the singularities of Legendre transforms, of equidistants, of derivatives, and of primitive hypersurfaces, coincide. Hence it suffices to study any of these objects to master all of them.

Riemannian geometry is a special case of the symplectic geometry of pairs of hypersurfaces in symplectic manifolds. This remark allows one to use the experience and intuition of Riemannian geometry and even of the elementary geometry of surfaces in the usual Euclidean 3-space to derive results in symplectic and contact geometries of pairs of hypersurfaces.

The resulting theory has useful applications to other problems in 3-space, in particular to variational problems with one-sided constraints, holonomic or not, and to optimal control problems.

The power of symplectic and contact geometries depends on the unification of many apparently disjoint branches of mathematics that these theories provide. It is comparable to the unification of most branches of mathematics provided by linear algebra or, more geometrically by projective geometry.

Remark. In the 19th Century there was a strange discussion between the mathematicians about which science is more important: Euclidean geometry or projective geometry?

The last word was said by Arthur Cayley who claimed:
"*Projective geometry is all geometry*", having in mind the interpretation of Euclidean geometry as of the study of geometrical objects in the projective space equipped with some additional structure at infinity, providing the Euclidean structure to the affine space (see Klein ?).

Today, projective geometry is accepted as the most basic one. Nowadays one could say "symplectic geometry is all geometry", but we formulate it in a more geometric way: *Contact geometry is all geometry.*

The difference between projective geometry and contact geometry is similar to the difference between algebra and calculus: projective geometry studies some finite-dimensional objects and groups, and contact geometry provides their infinite-dimensional versions.

Chapter 17

Elliptic coordinates and confocal hypersurfaces

Mechanics is the art of clever choice of coordinates *

Let $A : \mathbb{R}^n \to \mathbb{R}^{n*}$ be a symmetric operator in Euclidean space. Consider the following one-parameter family ("pencil") of such linear operators:

$$A_\lambda = A - \lambda I \ , \quad \lambda \in \mathbb{R} \ ,$$

where the operator I, which defines the Euclidean structure, is the "identity operator" for the natural identification of Euclidean space with its dual space.

We shall mostly consider the generic case where all the eigenvalues of the operator A are pairwise different.

Definition. The *confocal family* of linear operators is the family of the dual "resolvent" operators

$$A_\lambda^{-1} = (A - \lambda I)^{-1} : \mathbb{R}^{n*} \longrightarrow \mathbb{R}^n \ .$$

We associate to the initial family of linear operators the family of quadratic forms

$$F_\lambda(x) = \frac{1}{2}\big((A - \lambda I)x, x\big) \ , \quad x \in \mathbb{R}^n \ ,$$

and the following family of quadric hypersurfaces in Euclidean n-space defined by these forms:

$$M_\lambda = \{x \in \mathbb{R}^n : F_\lambda(x) = 1\} \ .$$

*A. I. Ishlinsky, quoted in the book on the Russian ballistic rockets [54], p. 70.

Similarly, to the family of dual operators we associate the family of quadratic forms that consists of the *confocal forms*

$$G_\lambda(y) = \frac{1}{2}\big((A-\lambda I)^{-1}y, y\big) , \quad y \in \mathbb{R}^{n*} .$$

The family of the *(dual) confocal quadric hypersurfaces* in the dual Euclidean n-space consists of the quadrics,

$$N_\lambda = \{y \in \mathbb{R}^{n*} : G_\lambda(y) = 1\} .$$

Remark. The hypersurface N_λ is dual to hypersurface M_λ in the sense that every point y of N_λ may be interpreted as an affine tangent space $\{x : (y,x) = 1\}$ to the hypersurface M_λ. And vice versa, the points of M_λ represent the tangent affine spaces to N_λ.

The quadratic form G_λ is the Legendre transform of F_λ.

Example. Consider a positive definite matrix, which determines an operator, in the Cartesian orthonormal coordinates u, v of the Euclidean plane,

$$\frac{1}{2}A = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} .$$

The corresponding quadratic form is

$$F_\lambda = (a-\lambda)u^2 + (b-\lambda)v^2 .$$

The hypersurface M_λ is the conic section $F_\lambda = 1$. It is an ellipse or hyperbola, depending on the situation of the parameter value λ with respect to numbers a and b.

The dual form is

$$G_\lambda = \frac{U^2}{a-\lambda} + \frac{V^2}{b-\lambda} ,$$

where U and V are the dual orthonormal coordinates in the dual Euclidean plane $(U|u = 1,\ U|v = 0,\ V|u = 0,\ V|v = 1)$ – Fig. 17.1.

Theorem 1. *For all different values of the parameter λ the conic sections $N_\lambda = \{(U,V) : G_\lambda(U,V) = 1\}$ have the same focal points.*

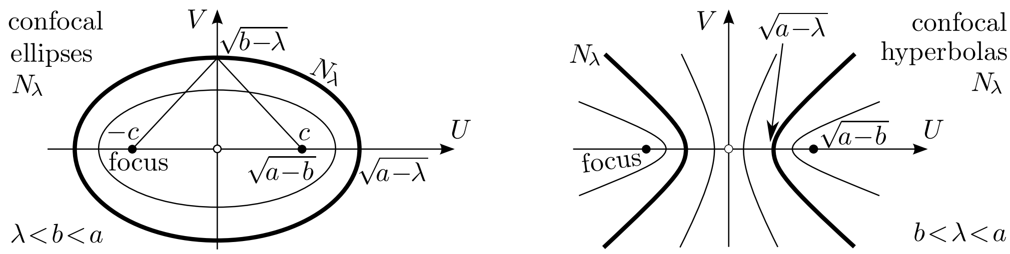


Figure 17.1: Families of confocal ellipses N_λ and of confocal hyperbolas N_λ with the same focal points.

Proof. Suppose, for simplicity, that $a > b > \lambda$. The large semi-axis of the ellipse is $\sqrt{a-\lambda}$, the small semi-axis is $\sqrt{b-\lambda}$, the focal points are on the U axis at distance c from the origin, where $c^2 = (a-\lambda) - (b-\lambda) = a - b$ does not depend on λ.

In the case $a > \lambda > b$, we obtain similarly that the focal points of the hyperbola, $(U = \pm\sqrt{a-b}, V = 0)$, do not depend on λ. $\square$

Theorem 2. *An ellipse and a hyperbola with coinciding focal points intersect orthogonally.*

Proof. (Fig. 17.2)

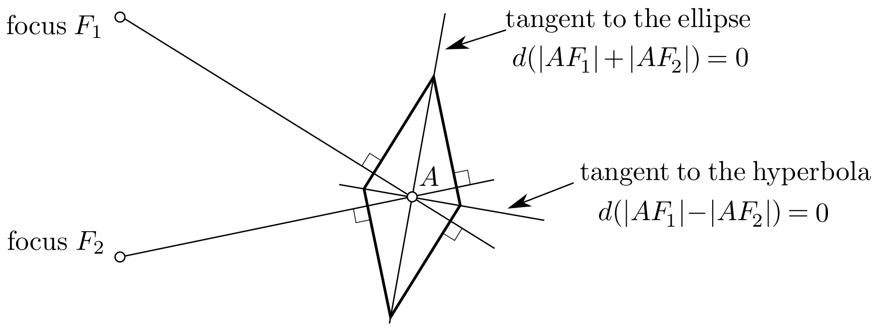


Figure 17.2: The infinitesimal rhombus, proving the orthogonality of confocal ellipse and hyperbola at their common point A.

Along the ellipse the sum of the distances to the focal points does not change, while along the hyperbola the difference of those distances remains constant.

It follows that the directions of the hyperbola and of the ellipse at the intersection point A are orthogonal. Indeed, construct the four lines (circular arcs)

$$\ell_1^{\pm} = \{\widetilde{A} : |F_1\widetilde{A}| = |F_1A| \pm \varepsilon\} , \quad \ell_2^{\pm} = \{\widetilde{A} : |F_2\widetilde{A}| = |F_2A| \pm \varepsilon\} .$$

These four lines form an infinitesimal rhombus (since both heights are of length 2ε).

The two diagonals of the rhombus are orthogonal. One diagonal is tangent to the ellipse, when the intersecting lines of $\ell_1^{\pm}$ and of $\ell_2^{\pm}$ correspond to opposite signs of ε, and the other to the hyperbola, for equal signs of ε. Theorem 2 is thus proved. □

Theorem 3. *Any point $y \in \mathbb{R}^{n*}$ belongs to n different quadric hypersurfaces $(N_{\lambda_1}, \ldots, N_{\lambda_n})$ of a given confocal family in Euclidean n-space and these hypersurfaces are pairwise orthogonal at that intersection point y.*

Definition. The numbers $(\lambda_1, \ldots, \lambda_n)$ are called *the elliptic coordinates* of the point y.

The above remarkable orthogonality property of the confocal hypersurfaces N_λ is projectively dual to the orthogonality of the principal directions of a linear family of quadratic forms F_λ. Now we shall prove Theorem 3.

The point $y \in \mathbb{R}^{n*}$ may be considered as a linear function $y : \mathbb{R}^n \to \mathbb{R}$. In dual terms, the fact that this point belongs to the quadric hypersurfaces N_{λ_k} means that the hyperplane $\Pi = \{x : y(x) = 1\}$ in $\mathbb{R}^n$ is tangent to the quadric hypersurface M_{λ_k} at a point $x_k \in M_{\lambda_k}$ for each of the n hypersurfaces M_{λ_k}.

To find the tangency points of the hyperplane $\Pi = \{x \in \mathbb{R}^n : (y, x) = 1\}$ with the quadrics M_λ (where $F_\lambda(x) = \frac{1}{2}\big((A - \lambda I)x, x\big) = 1$), consider the quadratic form

$$\frac{1}{2}(Bx, x) = \frac{1}{2}(Ax, x) - (y, x)^2 , \quad x \in \mathbb{R}^n .$$

Geometric Lemma. *The tangency points of the hyperplane Π with the quadrics M_λ are exactly the eigenvectors of the symmetric operator B that belong to the hyperplane Π.*

Proof. On the one hand, an eigenvector x of the form B verifies the conditions $Ax - 2(y, x)y = \lambda x$, that is,

$$Ax - \lambda x = 2y, \qquad \text{for } x \in V. \tag{1}$$

On the other hand, a tangency point x of the hyperplane Π with M_λ verifies the equations $(\text{grad } F_\lambda)(x) = cy$ for some $c \in \mathbb{R}$, that is,

$$Ax - \lambda x = cy \ , \quad F_\lambda(x) = 1 \ , \quad x \in V \ . \tag{2}$$

To find the value of c we use (2) and the fact that $F_\lambda = 1$ at the tangency point x:

$$1 = F_\lambda(x) = \frac{1}{2}(Ax - \lambda x, x) = \frac{c}{2}(y, x) = \frac{c}{2} \ .$$

That is, $c = 2$ at the tangency point x of the quadric M_λ with the hyperplane Π. Consequently, the tangency condition (2) implies the eigenvector condition (1) and vice versa. The Geometric Lemma is thus algebraically proved. □

Thus, the n orthogonal eigendirections of the quadratic form B are provided by the vectors from O to the points x_k at which the hyperplane Π is tangent to the n respective quadrics M_{λ_k} of the family – Fig. 17.3.

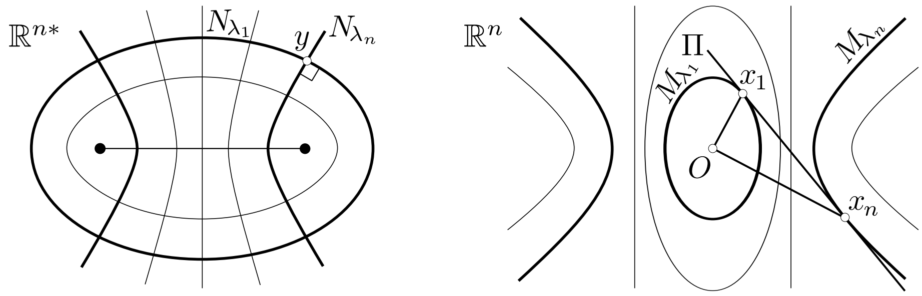


Figure 17.3: The orthogonality of the n confocal hypersurfaces N_{λ_k}, at the point y of $\mathbb{R}^{n*}$, is dual to the orthogonality of the eigenvectors x_k of the operator B in $\mathbb{R}^n$ at which the hyperplane Π is tangent to the quadrics M_{λ_k}.

For the dual confocal family, we get the n quadrics N_{λ_k} that contain the point y of $\mathbb{R}^{n*}$. Their orthogonality means (by the duality) the pairwise orthogonality of the eigenvectors x_k of the operator B in $\mathbb{R}^n$. Theorem 3 is thus algebraically proved.

Theorem 4 (Chasles)**.** *Given a family of confocal quadrics in n-dimensional Euclidean space, any generic straight line is tangent to $n-1$ quadrics of the family. The tangent spaces to the quadrics at the points of tangency with the line intersect pairwise orthogonally along that line.*

Proof. We project the quadrics of the family along the straight lines parallel to the given line, to the $(n-1)$-dimensional subspace $\mathbb{R}^{n-1*}$ orthogonal to those lines of Euclidean space $\mathbb{R}^{n*}$.

Each quadric defines an *apparent contour* in $\mathbb{R}^{n-1*}$, which by definition is the set of critical values of the projection of the quadric to $\mathbb{R}^{n-1*}$. Each apparent contour is itself a quadric, that is, a degree 2 hypersurface* in the $(n-1)$-dimensional orthogonal space $\mathbb{R}^{n-1*}$.

Lemma. *The apparent contours of the quadrics of a confocal family form themselves a confocal family of quadrics.*

Proof. The duality transforms sections to projections and vice versa. The apparent contours of the projections of confocal quadrics along the family of parallel lines are therefore dual to the sections of the dual quadrics by a hyperplane, containing the origin.

The sections of the quadrics of a Euclidean pencil form a Euclidean pencil of quadrics in the intersecting hyperplane.

By duality the dual quadrics form a confocal family, being just the initial apparent contours. □

Returning to the proof of Theorem 4, consider the orthogonal projection to $\mathbb{R}^{n-1*}$ along the lines parallel to the given straight line.

Since the $n-1$ apparent contours are confocal quadrics in $\mathbb{R}^{n-1*}$, they intersect pairwise orthogonally at the generic points z of $\mathbb{R}^{n-1*}$.

The $n-1$ confocal quadrics of the initial family in $\mathbb{R}^{n*}$, whose apparent contours contain the point z, are tangent to the straight line Z projected to z.

---

*The smoothness of the visible contour of a quadric is a peculiar property: for the projection of a generic smooth surface, its visible contour has singularities. Algebraic objects of low degree are not generic; the singularities would appear for the projection of a generically perturbed non convex quadric.

The absence of these singularities for the projections of the quadrics is similar to the absence of curvature for the degree 1 hypersurfaces.

The general theorems of singularity theory imply that the algebraic objects become locally generic when the degree is sufficiently high. The asymptotic lines on the hyperbolic quadrics are non generic (being closed curves). It is conjectured that they behave globally in a generic way on a cubic surface (where they might even fill densely some domains, due to the absence of the analytic first integrals, similar to the situation in the 3-body problem).

The orthogonality of the intersecting apparent contours in $\mathbb{R}^{n-1*}$ means the orthogonality of the tangent spaces to the initial quadrics at the $n-1$ points of tangency with the projecting straight line Z.

Theorem 4 is proved. □

Theorem 5 (Jacobi and Chasles)**.** *Given a geodesic line on a quadric Q in an n-dimensional Euclidean space, there exist $n-2$ quadrics confocal to Q, such that all the tangent lines to the geodesic are also tangent to these $n-2$ quadrics.*

Proof. Consider the manifold of the oriented straight lines in Euclidean n-space. This manifold has a natural symplectic structure as the manifold of the characteristics in the constant energy hypersurface $p^2 = 1$ of the phase space of a free particle, moving under its own inertia in Euclidean space.

The *characteristics* on a hypersurface of a symplectic manifold are the integral curves of the field of directions that are symplectically orthogonal to the tangent hyperplanes of the hypersurface (called *characteristic directions*). In other words, they are the phase curves of the Hamilton differential equations whose Hamilton function vanishes (to the first order) on the hypersurface.

The symplectic structure $\widetilde{\omega}$ of the manifold of the characteristics on a hypersurface of a symplectic manifold is defined by the condition that its value on any two tangent vectors of the manifold of the characteristics at the same point is equal to the value of the initial symplectic structure ω of the ambient symplectic manifold on any two tangent vectors to the given hypersurface at the same point that are projected to the given two tangent vectors of the characteristic manifold (under the natural projection of the hypersurface onto its manifold of the characteristics).

The fact that the value of $\widetilde{\omega}$ is independent of the choice of these two tangent vectors (with given projections) follows from Stokes Lemma.

Lemma A. *Each characteristic of the manifold of the straight lines tangent to a given hypersurface in Euclidean space, consists of the lines tangent to a single geodesic line of the hypersurface.*

Proof of Lemma A. To simplify the notations, we will identify the cotangent vectors of Euclidean space with the tangent vectors by using the Euclidean structure and associating to a vector the 1-form that represents the scalar product with this vector: $v \leftrightarrow \langle v\, , \cdot \rangle$.

In this terminology the initial Euclidean space phase space is the tangent bundle formed by the velocities of the motions in Euclidean space.

The unit vectors tangent to the given hypersurface in Euclidean space form a smooth submanifold of codimension 3 in this phase space. Its characteristics are the geodesic lines of the hypersurface (considered together with the velocity vectors of the motion with velocity one along the geodesic line).

Associate to each tangent vector of Euclidean space the straight line containing it. The codimension 3 submanifold of the of the manifold of the unit tangent vectors of the hypersurface is sent by this map φ onto the manifold of the tangent straight lines of the hypersurface.

For the symplectic structure defined above of this manifold of straight lines, φ sends the characteristics to the characteristics. Lemma A follows. $\square$

Suppose now that f is a smooth function on Euclidean space, whose restriction to some straight line has a non degenerate critical point. Then on any neighbouring straight line the function f has a neighbouring critical point. The corresponding critical values determine a function F on the manifold of the straight lines. We shall say that the function F on the manifold of the straight lines is induced by the initial function f.

Lemma B. *Suppose that a straight line Z in Euclidean space is tangent to two smooth hypersurfaces $f = \mathrm{const}$ and $g = \mathrm{const}$ (at two points, eventually different). Suppose that the tangent hyperplanes of these hypersurfaces at these two points are orthogonal.*

Then the Poisson bracket of the two corresponding induced functions F and G on the space of the straight lines is zero at the common tangent line Z.

Proof of Lemma B. We shall calculate the derivative of the function G along the phase flow of the Hamilton function F. The phase curves of F that belong to the level hypersurface $F = \mathrm{const}$ are the characteristics of this hypersurface (in the symplectic manifold of the straight lines).

The level hypersurface of F consists of those straight lines that are tangent to a single level surface of the function f in Euclidean space. By Lemma A, each characteristic of this hypersurface consists of the straight lines tangent to a fixed geodesic line of the level surface of the function f.

For an infinitesimally small displacement of a point along the geodesic in a surface, the tangent straight line of the geodesic curve rotates (up to infinitesimal quantities of higher order) in the 2-plane spanned by the original tangent straight line and the normal line to the hypersurface.

By the Lemma condition, the tangent plane to the level surface of the function g (at the tangency point with the line Z) is orthogonal to the tangent plane of the level surface of the function f (at the initial point of the line Z).

Therefore, under the above-mentioned infinitesimally small rotation, the straight line remains tangent to the same level surface of the function g (up to a higher order infinitesimal). It means that the rate of change of the function G along the phase flow of F is equal to zero. Lemma B is proved. □

End of the proof of Theorem 5. We fix a generic straight line in Euclidean space $\mathbb{R}^n$. According to Theorem 4, this line is tangent to $n-1$ quadrics of the confocal family, at $n-1$ points. We construct in the neighbourhood of each of these points a smooth function without critical points, whose level surfaces are the quadrics of our confocal family.

We fix one of these quadrics and call it "the first quadric". Consider the Hamilton system on the manifold of the straight lines, whose Hamilton function F is induced by the first of the constructed functions, f.

The remaining induced functions have zero Poisson brackets with F by Lemma B (the orthogonality of the tangent spaces is provided by Theorem 4).

Thus all the $n-1$ induced functions F_k are first integrals of the Hamilton system, generated by any of them.

Since the straight lines tangent to a geodesic line of the first quadric form a phase curve of the first Hamilton system (Lemma A), all the induced functions take constant values along this phase curve. This constancy statement proves Theorem 5 □

The above proof provides also the following complete integrability result:

Theorem 6. *The geodesic flow on a central surface Q of degree* 2 *in Euclidean space is a completely integrable Hamilton system, having as many first integrals in involution as it has degrees of freedom.*

In the notations of the preceding proof, these integrals are the $n-1$ induced functions F_k. In other words, the integrals are the elliptic coordinates λ_k of the quadrics N_{λ_k} of the confocal family of the quadric Q, to which the tangent straight lines of a geodesic line of Q are all tangent (according to Theorem 5).

The elliptic coordinate hypersurfaces are the quadrics N_λ. Choosing the Cartesian coordinates $y = (y_1, \dots, y_n)$ in Euclidean space $\mathbb{R}^n$, we can reduce the symmetric operator A and the quadratic forms F_λ and G_λ to the diagonal forms. Suppose the eigenvalues of the operator A are all different and denote them by

$$a_1 < a_2 < \cdots < a_n.$$

In these coordinates, the equation of the quadric N_λ is $G_\lambda(y) = 1$, that is,

$$\frac{1}{2}\left(\frac{y_1^2}{a_1 - \lambda} + \cdots + \frac{y_n^2}{a_n - \lambda}\right) = 1\ .$$

The topological type of this quadric depends on the position of the number λ with respect to the eigenvalues – Fig. 17.4.

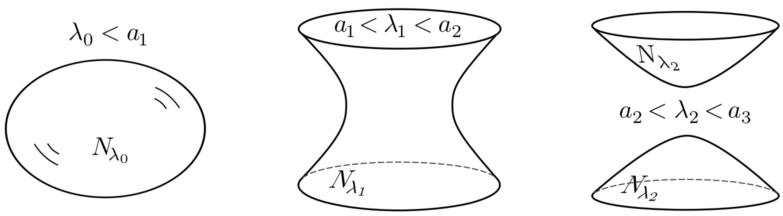


Figure 17.4: The evolution of the quadric N_λ of a confocal family.

If $a_k < \lambda < a_{k+1}$ the quadratic form G_λ has exactly k squares with the minus sign.

For $n = 3$ the quadric N_{λ_0} is an ellipsoid for $\lambda_0 < a_1$, degenerating for $\lambda = a_1$ (flattening to become a two-sided ellipse plane surface). For $a_1 < \lambda_1 < a_2$ the confocal surface N_{λ_1} is a one-sheet hyperboloid, diffeomorphic to the product $S^1 \times \mathbb{R}$. It degenerates once more for $\lambda = a_2$ (flattening to become a two-sided connected plane surface, bounded by a hyperbola).

Next confocal surfaces N_{λ_2} ($a_2 < \lambda_2 < a_3$) are two-sheets hyperboloids (diffeomorphic to the product $S^0 \times \mathbb{R}^2$). For $\lambda = a_3$ this hyperboloid degenerates the last time, disappearing at infinity and being void for $\lambda > a_3$.

In the n-dimensional case, the quadric N_λ is diffeomorphic to the product $S^{n-1-k} \times \mathbb{R}^k$ for $a_k < \lambda < a_{k+1}$. Therefore, the signatures of the n confocal quadrics that contain a given (generic) point cover the n possible different values $0 \leqslant k < n$.

17.1 Gravitational and magnetic fields on hyperboloids

The theory of elliptic coordinates extends to the hyperboloids the classical Newton theory of the ellipsoids attraction. The gravitational or electrostatic fields of this Newtonian theory should be replaced in these hyperbolic generalisations by the electromagnetic field in the 3-dimensional case ($n = 3$) and by its higher-dimensional generalisations for $n > 3$.

We start from the Newton-Ivory ellipsoids attraction theory.

Definition. A *homeoidal density* on the surface of an ellipsoid E is the density of a layer between E and an infinitely nearby ellipsoid with the same centre, homothetic to E (the layer being filled by a constant density gravitating mass).

Theorem 7 (Newton–Ivory)**.** *A finite mass distributed on the surface of an ellipsoid with homeoidal density, does not attract any internal point. It attracts every external point in the same way as if the gravitating mass were distributed with homeoidal density on the surface of a smaller confocal ellipsoid.*

The attraction is determined here by the law of Newton or that of Coulomb: In the n-dimensional Euclidean space the force is proportional to r^{1-n} at distance r, as it is prescribed by the fundamental solution of the Laplace equation, $\Delta U = \delta$ for the force grad U.

The spherical ellipsoid is already interesting: In this case the non-attraction of the internal points was discovered by Newton, calculating the compensation of the attractions by the opposite points, while the external case corresponds to the possibility of the replacement of the spherically symmetric mass distributions by an attracting point concentrating all the mass at the centre, which is crucial for the Newton theory of the planetary attraction.

The Newton theorem on the non-attraction of the internal points is a particular case of the following general result, which replace the quadrics by the algebraic hypersurfaces of arbitrary degree.

Definition. A polynomial $f(x_1, \ldots, x_n)$ of degree m is *hyperbolic*, if its restriction to every real straight line that contains the origin has all its m roots real.

Remark. More generally, an algebraic hypersurface of degree d in the real projective space $\mathbb{R}P^n$ is called *hyperbolic* with respect to a point t outside this hypersurface, called the "time direction", if every real projective straight line that contains t intersects the hypersurface at d real points. The hypersurface is called *strictly hyperbolic*, if all these d intersection points are different. A strictly hyperbolic hypersurface consists of several connected components, diffeomorphic to the spheres if d is even. The components are numbered by the order of their points on each ray through the point t.

Example. The cone formed by the degenerate quadratic forms in $\mathbb{R}^m$ determines a hyperbolic hypersurface in the projective version $\mathbb{R}\mathrm{P}^n$ of the space of the real quadratic forms, $\mathbb{R}^n$ with $n = m(m+1)/2 - 1$.

This hypersurface is hyperbolic with respect to any direction t of a positive definite form, since each quadratic form in Euclidean m-space has m real principal axis – the characteristic equation is of degree $d = m$. This hyperbolic hypersurface is strictly hyperbolic only for $m \leqslant 2$.

The theory of the elliptic coordinates is based on the hyperbolicity of the hypersurface of the degenerate quadratic forms. This hyperbolicity statement means that all the m eigenvalues of a form in $\mathbb{R}^m$ are real.

It would be interesting to understand whether the theory of the elliptic coordinates could be generalised to other hyperbolic hypersurfaces.

The first natural examples are the hypersurfaces of the degenerate Hermitian forms and of the degenerate hyper-Hermitian forms, which both are hyperbolic with respect to the positive definite directions.

Such generalisations of the hyperbolic hypersurfaces to other compact Lie algebras would be also interesting, but even in the case of the Hermitian forms the elliptic coordinates theory is yet missing (hopefully, temporarily).

Definition. A *homeoidal charge on the zero hypersurface* $f = 0$ of a hyperbolic polynomial is defined as the density of a homogeneous infinitely thin layer between the hypersurfaces $f = 0$ and $f = \varepsilon$ $(\varepsilon \to 0)$, for which the signs of the charges are chosen in such a way that the successive ovaloids have opposite charge signs.

Theorem 8. *A homeoidal charge does not attract the origin (nor any other point within the innermost ovaloid).*

This property is preserved if the homeoidal charge density is multiplied by any polynomial of degree at most $m - 2$.

When the homeoidal density is multiplied by any polynomial of degree $m-2+r$, the attracting potential of the resulting charge inside the innermost ovaloid becomes a harmonic polynomial of degree r (see [78])

Trying to extend to the hyperboloids the Ivory theorem on the confocal ellipsoids attraction, one sees that the topological difference between these hypersurfaces implies the necessity of replacing the homeoidal densities by harmonic differential forms of different degrees (depending on the signatures of the hyperboloids). The Newton or the Coulomb attraction laws should be replaced by the corresponding generalised form-potentials provided by the generalisations of the Biot-Savart law.

In the simplest nontrivial case of a hyperboloid of one sheet in three-dimensional Euclidean space, the result is the following theorem on magnetic fields.

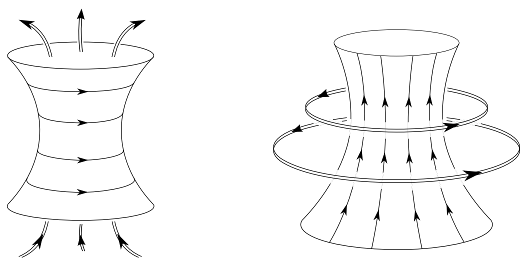


Figure 17.5: Magnetic fields generalising the theorems of Newton and Ivory.

The hyperboloid divides the space into two parts: "internal" and "external", the latter being non-simply connected (Fig. 17.5). The lines of the elliptic coordinate system on our hyperboloid, which are obtained by intersecting it with the confocal ellipsoids, are the closed lines of curvature of the hyperboloid surface. They are called the *parallels* of the hyperboloid.

The curves obtained by intersection with the confocal hyperboloids of two sheets are orthogonal to the parallels and are called the *meridians* of the hyperboloid surface.

Although the elliptic coordinate system has singularities (on each symmetry plane of the quadrics of the family), the hyperboloid surface is smoothly fibred into its parallels (diffeomorphic to the circle) and also into its meridians (diffeomorphic to the line) – see Fig. 17.5.

The region inside the hyperboloidal tube is smoothly fibred by the internal meridians orthogonal to the confocal ellipsoids of the family. The annular region outside the hyperboloidal tube is smoothly fibred by the closed external parallels orthogonal to the two-sheet confocal hyperboloids of the family. Both fibrations of "magnetic lines" are shown in Fig. 17.5 by double lines.

Theorem 9. *A current with a suitable density directed along the meridians of a hyperboloid produces a magnetic field that is zero inside the hyperboloidal tube, while outside it the field is directed along the external parallels.*

A current with a suitable density directed along the parallels of a hyperboloid produces a magnetic field that is zero outside the hyperboloidal tube, but inside it the field is directed along the internal meridians – Fig. 17.5.

The current densities that give rise to such magnetic fields generalise the homeoidal densities on the ellipsoids. They can be described in the following way.

To each family of confocal quadrics in 3-dimensional Euclidean space are associated two "focal curves": An ellipse and a hyperbola – Fig. 17.6.

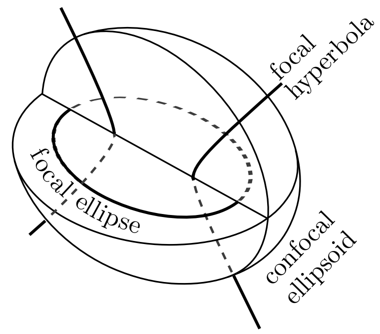


Figure 17.6: Focal ellipse and focal hyperbola of a family of confocal quadrics.

The focal ellipse is the boundary of the limiting ellipsoid of the family, whose shortest axis shrinks to zero. The focal hyperbola arises in a similar way from the transition between the one-sheet hyperboloids and the two-sheet hyperboloids of the family of confocal quadrics.

The planes of the focal ellipse and of the focal hyperbola are orthogonal. Each of the two focal curves contains the focal points of the other focal curve.

The homeoidal density on the focal ellipse is constructed in the following way. Consider the infinitesimally thin parallel "wire" that is the intersection of the following two layers: The layer between the initial ellipsoid and a ε-homothetic one and the layer between the initial one-sheeted hyperboloid and a ε-homothetic to it – both homotheties are taken with respect to the common centre of the confocal family. We normalise this homeoidal density on the parallel wire in such a way that the total mass of the whole parallel wire is equal to one.

The normalised homeoidal densities on these parallels converge to some limiting density on the focal ellipse when these parallels approach the focal ellipse.

The homeoidal density on the focal hyperbola is defined by a similar construction, but starting from the "meridian wires".

Now we describe the "suitable" currents of Theorem 9. The surface of the one-sheet hyperboloid is fibred over the focal ellipse (the fibre over a point is the meridian that lies on the same confocal hyperboloid with two sheets as that point).

The flux of the "suitable current along the meridians" of Theorem 9 through any curve on the one-sheet hyperboloid is equal to the integral of the homeoidal density on the focal ellipse over the projection of that curve to the focal ellipse (along the confocal hyperboloids of two sheets).

The density of the flow along the parallels is induced in an analogous way from the homeoidal density on the focal hyperbola.

More details on these applications of the elliptic coordinates are published in [17, 18, 19, 20, 117].

Inside the hyperboloidal tube, the magnetic field induced by the current directed along the parallels (of Theorem 9) coincides, up to the sign, with the Newtonian or the Coulombian field produced by a charge distributed with homeoidal density along a confocal ellipsoid (in the part of the tube external to this ellipsoid).

This is the density with which a charge will distribute itself along the surface of a conductive "metallic" ellipsoid.

In the annular domain external to the hyperboloidal tube, the magnetic field produced by the current directed along the meridians (of Theorem 9) coincides, up to the sign, with the Coulombian field produced by two equal charges with opposite signs, distributed with the homeoidal density on the two sheets of the confocal two-sheet hyperboloid. The coincidence takes place along the part of the external annular domain situated between the two sheets of this confocal hyperboloid. This result is due to O. P. Shcherbak.

The extension of all this 3-dimensional magnetic theory to the Euclidean spaces of higher dimension is due to B. Z. Shapiro and A. D. Vainstein.

For a hyperboloid in $\mathbb{R}^n$, diffeomorphic to $\mathbb{S}^k \times \mathbb{R}^\ell$, one constructs a harmonic k-form on the "exterior region" (diffeomorphic to the product of S^k with a half-space) and a harmonic ℓ-form on the "interior domain".

The corresponding homeoidal densities are defined on the focal ellipsoid of codimension k and on the focal hyperboloid of two sheets of codimension ℓ, by a limit procedure, similar to that described above for $k = \ell = 1$ (using the intersections of two layers between infinitesimally close and homothetic quadrics).

Unfortunately, in spite of the geometric content of the statements of these generalised versions of Theorem 8, their non computational geometric proofs are still missing, even in the special case of the magnetic fields in Euclidean 3-space, discussed in Theorem 9.

Remark. These results provide some distinguished harmonic forms on the hyperboloids in the Euclidean spaces and on their complementary domains.

Unfortunately, it is not clear how to choose such distinguished representatives in the spaces of the harmonic differential forms on the non compact algebraic manifolds of higher degree or on their complements.

It should be some partial differential equations version of the mixed Hodge structures of algebraic geometry, but it is missing in the literature.

Other interesting extensions of the theory of elliptic coordinates was proposed in Problem 1981–29 of [36]. See also some attempts to answer these problems in [127].

Namely, one might extend the elliptic coordinates to the infinite-dimensional Hilbert spaces. Both cases are interesting: That of the discrete spectre operators and that of the continuous spectre.

In the case of the discrete spectre, one should replace the sums by the series and to prove the convergence those series for suitable asymptotic behaviour of the discrete spectre.

In the case of the continuous spectre, the sums should be replaced by the integrals and the convergence problems are more difficult.

It may be useful to observe that some continuous spectre elliptic coordinates are provided by the so-called Hilbert transforms. Strangely, Hilbert did not pointed out the relations of his formulae to the Jacobi formulae that express the Cartesian coordinates of a point in terms of its elliptic coordinates.

The Jacobi theory contains also a remarkable formula, which represents the Riemannian metric on the quadric in terms of the elliptic coordinates.

The resulting expressions for the momenta of this metrical tensor are strangely dual to the Jacobi expressions of the Cartesian coordinates of a point of Euclidean space in terms of its elliptic coordinates.

Consider the elliptic coordinates $(\lambda_1, \ldots, \lambda_n)$ defined in terms of the Cartesian coordinates $(x_1, \ldots, x_n)$ as the roots of the equation

$$\sum_{i=1}^{n} \frac{x_i^2}{a_i + \lambda} = 1.$$

The first of the two strangely dual formulae of Jacobi is

$$x_i^2 = \frac{\prod_j (a_i + \lambda_j)}{\prod_{j \neq i} (a_i - a_j)}. \tag{1}$$

Consider the metric form

$$4 \sum (dx_i)^2 = \sum M_j (d\lambda_j)^2.$$

The second formula of Jacobi provides the expressions of the inverted moments

$$M_j^{-1} = \frac{\prod_j (a_i + \lambda_j)}{\prod_{j \neq i} (\lambda_i - \lambda_j)}. \tag{2}$$

The explanation of the strange Jacobi duality between formulae (1) and (2) should interpret the squared lengths a_i of the semi-axes of the initial ellipsoid as the elliptic coordinates in some mysterious Euclidean space with Cartesian coordinates $M_j^{-1/2}$, while the numbers λ_j should be the squared lengths of the semi-axes of some ellipsoid "naturally defined by x" in this mysterious Euclidean space.

One can consider the theory of elliptic coordinates as the theory of two quadratic forms in two dual spaces: One in the first space and the second in the dual space. The Jacobi strange duality might reflect the involution that exchanges the two dual spaces and the roles of the two forms.

It would be interesting to compare this involution with the contactomorphism between the manifolds $J^1(\mathbb{S}^{n-1}, \mathbb{R})$ and $ST^*\mathbb{R}^n$, described on p. 603-616 – Fig. 16.19.

This strange algebraic duality has never been mentioned nor explained geometrically in the expositions of the theory of elliptic coordinates.

Finally, the complete integrability of the equations of the geodesics on the quadrics, provided by the theory of elliptic coordinates, should have infinite-dimensional extensions, which might be new and interesting completely integrable systems of ordinary or partial differential equations of mathematical physics.

One of the goals of including the theory of elliptic coordinates in the present elementary textbook was to approach the answers to the following four unsolved questions, already discussed above:

– infinite-dimensional elliptic coordinates in the discrete and continuous spectrum cases;

– Hilbert transforms as generalised elliptic coordinates;

– Geometric explanation of the duality of the Jacobi formulae;

– applications of infinite-dimensional elliptic coordinates to mathematical physics.

Even the finite dimensional version has interesting unusual applications in mathematical physics, like the following application to celestial mechanics of the Earth satellites.

Using elliptic coordinates, Jacobi solved the gravitational Newton equations of a point attracted by two stationary attracting points of equal masses, say, at $z = \pm\varepsilon$.

This explicit solution could be used to approximate the motion of a mass point in the gravitational field created by a planet if this planet would be “lemon-shaped” elongated: The distance from its centre to the poles being ε-greater than the radius of the equator.

Unfortunately, the real earth is not “lemon-shaped” but “tomato-shaped”: The distance from its centre to its poles is smaller than the equatorial radius by approximately 1/300 of its length (about 20 kilometres).

The unusual application of the elliptic coordinates to this tomato type case is the following trick: one replaces the two real attracting points $\{z = \pm\varepsilon\}$ by a similar system of two imaginary attracting points $\{z = \pm\sqrt{-1}\varepsilon\}$. The Jacobi formulae are still applicable. This trick reduces the tomato-type case to the lemon-type case. It is similar to the interpretation of the Lobachevsky plane as of a sphere of imaginary radius, and it is extremely useful for the practical space-researches of the satellites orbits.

Chapter 18

Vassiliev theory of knots and discriminants

In these times, the angel of topology and the devil
of abstract algebra fight for the soul
of each individual mathematical domain
H. Weyl*

The study of the discriminant variety in a functional space of smooth maps is a traditional and fundamental part of the theory of singularities. The discriminant variety is the set of those points of the functional space that represent the maps having non generic singularities. The topologic, homotopic and even homologic invariants of the complement to the discriminant variety (that is, of the space of generic maps) are important for many applications. However, the progress in these difficult global problems of singularity theory was rather slow until Vassiliev [132], in the beginning of the 90's, showed the new perspectives opened up by the singularity theory approach in knot theory.

18.1 Space of knots and discriminant

A *knot* is a connected component in the space of smooth embeddings of a circle into 3-space. Hence, we start with the functional space $\mathcal{F}$ of all

---

* *Invariants*, Duke Math. J., 5 (1939). This description seems to be an allusion to a painting by Uccello (at the Urbino castle) "l'hostie profanée", representing an event that happened in Paris in 1290. The event is also represented in a series of pictures in the church Saint Jean-Saint François in Paris. In the best version, which was shown in the museum of Colmar, an angel and a devil are fighting for the soul of a tortured lady.

smooth maps of $\mathbb{S}^1$ into $\mathbb{R}^3$ and we define the *discriminant variety* Σ as the set of maps that are not embeddings (that is, those that have either self-intersections or singularities – see Fig. 18.1).

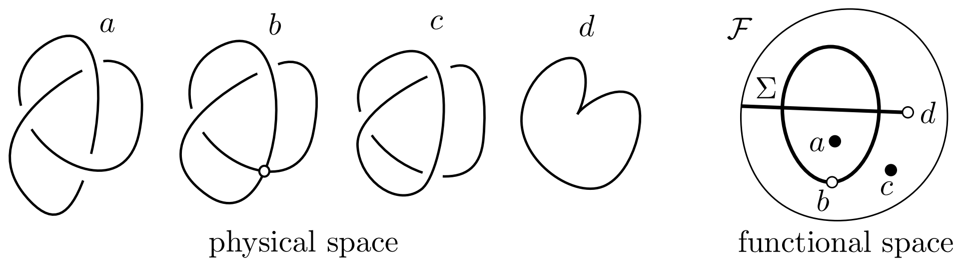


Figure 18.1: Generic maps and the discriminant variety in the space $\mathcal{F}$ of maps $\mathbb{S}^1 \to \mathbb{R}^3$.

The discriminant variety is a *hypersurface* in the space $\mathcal{F}$ of all maps, since the self-intersections occur in generic one-parameter families of maps of a curve in 3-space. This hypersurface subdivides the complement into connected domains, which are the knots.

We wish to study the topologic properties of the knot space $\mathcal{F} \setminus \Sigma$. For instance, the elements of its 0-dimensional cohomology group are locally constant functions, that is, knot invariants. Such functions can be multiplied, forming a ring:

$$H^0(\mathcal{F} \setminus \Sigma) = \text{ring of knot invariants} .$$

The space $\mathcal{F}$ is a vector space and, hence, is contractible. The study of the cohomology of the space of knots is therefore reducible to that of the discriminant, modulo ∞, by the Alexander duality *. The difficulties of the infinite dimensionality of $\mathcal{F}$ can be overcome by the standard finite dimensional approximation technique of singularity theory (see e.g., [8], [14], [16], [129], [39], [40], [133], [23], [131]). For instance, one can replace $\mathcal{F}$ by the space $\mathcal{F}_N$ of trigonometric polynomials of degree at most N. Then each homology group $H^i(\mathcal{F}_N - \Sigma)$ stabilises for $N \to \infty$:

$$H^i(\mathcal{F}_N \setminus \Sigma) \approx H^i(\mathcal{F} \setminus \Sigma) \qquad \text{for } N >> 1 .$$

Thus the Alexander duality is essentially used only in finite dimensional cases.

*MENTION SOME WORDS ABOUT THE ALEXANDER DUALITY

The advantage of the discriminant variety over its complementary knot space, which is our main object of study, is that this variety is naturally stratified according to the hierarchy of the singularities, while the knot space is smooth. Thus, to study homology, we need to cut the knot spaces into pieces, while for the discriminant variety the pieces are provided by the strata of the stratification. This stratification induces an additional structure in the homology of the discriminant, which survives also in the cohomology of the knot space, for instance, in the ring of its zero-dimensional cohomology. This chapter is an introduction to the study of the Vassiliev structure in the ring of knot invariants.

The works of J. Birman, X.S. Lin, D. Bar-Natan, and M. Kontsevich ([50], [97], [45], [46]) have shown that this Vassiliev structure is a fundamental general combinatorial mathematical object, related to the Jacobi identity, Yang-Baxter and Knizhnik-Zamolodchikov equations, the hierarchy of Feynman integrals of perturbative theory in the Chern-Simons action, the D. Zagier ζ-functions of several variables, and to the cohomology of the Lie algebra of Hamiltonian vector fields on infinite dimensional spaces.

18.2 The knot classification problem

We hope that the reader has seen many knots and does understand the difficult mathematical problem of the classification of the knots.

In fact, this mathematical problem had been first formulated explicitly in 1867, [123], by a physicist, Sir Thomson, lord Kelvin, who accepted the existence of atoms, but considered as monstrous the assumption of infinitely strong and infinitely rigid pieces of matter. After the Helmholtz's discovery of the law of vortex motion in a *perfect liquid* (free of viscosity), Thomson suggested the hypothesis that "all bodies are composed of vortex atoms in a perfect homogeneous liquid". His vortex atoms (which he also called "Helmholtz atoms") were small knots along which a wave propagate periodically. This permitted him to unify the atomic and the wave ideas about matter.

He suggested that the geometric and topological properties of the vortex knot and the properties of the wave propagating along it were responsible for the chemical peculiarities of the different elements.

He tried then to establish tables of atoms by classifying knots. So, he started to classify the knots by studying their plane projections (today called *knot diagrams*) with few self-intersections of the projected closed line.

Even to understand whether two such projections represent the same knot (that is, whether one closed space curve may be transformed continuously to the other, remaining free of self-intersections during all the deformation) is a difficult task: Such combinatorial problems are closed to the so-called *algorithmically unsolvable problems* (a celebrated example of an algorithmically unsolvable problem is the problem to recognise whether a given finite system of polynomial equations with integer coefficients has an integer solution).

Kelvin's attempt was interred by Mendeleev's discovery of the very simple, but until then unnoticed, arithmetic relations between the properties of the chemical elements and their atomic masses. The result of that great discovery is called today *Mendeleev periodic table of elements*. Mendeleev published his discovery at the end of 1869, but his theory was unanimously accepted in Occident only five years later, after the discovery of new chemical elements predicted by him.

Today knot theory is an important branch of mathematics with applications in physics, including quantum field theory. In a sense, Kelvin's idea (about the fundamental role of topology in the study of elementary particles) is reviving, though knots are usually substituted with invariants of manifolds of larger dimensions.

Knot invariants. In order to distinguish the knots, people invented the *knot invariants*: Such characteristics of the knots that are algorithmically computable and that take equal values on any two representations of the same knot. Usually, knot invariants are constructed by using the characteristics of the knot diagrams (i.e., of their projections on the plane).

An example is the *Alexander Polynomial* (the first knot polynomial), which associates a polynomial with integer coefficients to each knot type. To determine the Alexander polynomial of a knot, one takes, first, an oriented knot diagram of that knot with n crossings: Such knot diagram divides the plane into $n+2$ regions. Next, one takes an $(n+2)\times n$ matrix whose $n+2$ rows correspond to the regions, and its n columns to the n crossings. The possible values of the matrix entries are either 0, 1, -1, x, $-x$. The entry associated to a particular region and crossing is 0 if the region is not adjacent to the crossing. If the region is adjacent to the crossing, then the entry is determined by the location of the region at the crossing, viewed from the incoming undercrossing line:

on the right before the crossing $\mapsto\ 1$;
on the right after the crossing $\mapsto -1$;
on the left before the crossing $\mapsto -x$;
on the left after the crossing $\mapsto\ x$.

Finally, one removes two rows from the matrix, corresponding to adjacent regions, and computes the determinant of the obtained $n\times n$ matrix, whose value will differ by multiplication by $\pm x^n$, depending on the removed rows. Normalising the determinant, in order to have a positive constant term, one gets the Alexander Polynomial. It is independent of the chosen knot diagram and of the choices involved in this algorithm to compute it (indeed the definition of the Alexander polynomial is independent of this algorithm: It uses the first homology, with integer coefficients, of the complement of the knot).

The genus of a knot, discussed in page 115, is another example of a knot invariant.

Vassiliev invariants are special invariants of the knots, whose position in the space of all knot invariants is similar to the position of polynomials in the space of all continuous functions.

These invariants are closely related to such branches of mathematics as singularity theory, complex integration theory, graph theory, configuration spaces theory, Lie algebras and quantum field theory.

They represent a happy part of the almost uncomputable invariants of the theory of knots, but the general pessimistic opinion has been, just for this reason, that they form a too small part of the complicated world of the invariants, insufficient for the goal of distinguishing the different knots.

The Vassiliev invariants and the Poincaré conjecture are related by the following theorem ([62]):
The Poincaré conjecture would follow from the fact that the Vassiliev invariants of knots were distinguishing any two different knots.

The Poincaré conjecture states that:
Any closed simply-connected 3*-manifold is homeomorphic to the* 3*-sphere.*
The corresponding characterisation of the 2-sphere follows from the classification of surfaces. Starting from dimension 5 one should add, to the "simply-connected" condition $\pi_1(M^3) = 0$, the higher homotopy groups vanishing conditions $\pi_k(M^n) = 0$, for all $k < n$. In this case the manifold is homeomorphic to the sphere $\mathbb{S}^n$ ("Smale's theorem").
So, the mild dimensions ($n = 3$ and 4) remain the most difficult cases of the Poincaré problem. The Poincaré conjecture for the sphere $\mathbb{S}^3$ was proved recently by G. Perelman, who elaborated the previous suggestions of R. Hamilton.

This unexpected relation between Vassiliev invariants and the Poincaré topologic ideas is restoring the simplest things priority: In spite of their unsophisticated nature, Vassiliev invariants (invented only 24 years ago) are universal. One hopes that they contain all the invariants of the knots in the 3-sphere, in the sense that any invariant is a function of the simplest Vassiliev invariants.

18.3 Vassiliev invariants

These invariants form an increasing sequence of finite dimensional subspaces in the ring of knot invariants, similar to the sequence of spaces of polynomials of increasing degrees in the ring of power series. Together these finite dimensional subspaces form the sub-ring V of the Vassiliev invariants:

$$H^0(\mathcal{F} \setminus \Sigma) \supset V \supset \cdots \supset V_n \supset \cdots \supset V_1 \supset V_0 \ .$$

The subspace V_n, (or the subgroup, if we consider cohomology with integer coefficients) is called *the space (group) of the Vassiliev invariants of order* n. The product of invariants of orders m and n will be an invariant of order $m+n$.

The polynomials of degree at most n are defined by the condition $d^{n+1}p = 0$. The Vassiliev invariants of order n are defined similarly:

$$\nabla^{n+1} i = 0 \ , \quad i \in H^0(\mathcal{F} - \Sigma) \ ,$$

where the *jump operator* ∇, which replaces the derivative (and which, we shall see, is also similar to the residue), is defined by the following construction.

Lemma. *The discriminant hypersurface in* $\mathcal{F}$ *has a natural coorientation* – Fig. 18.2

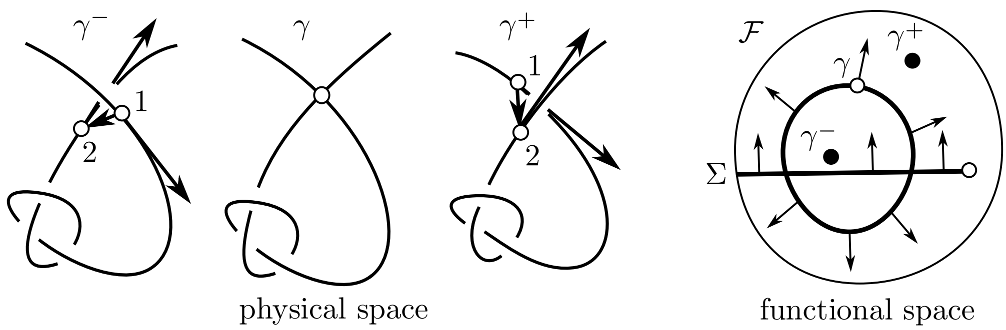


Figure 18.2: Coorientation of the discriminant variety.

Proof. Indeed, fix the orientations in the circle and in the 3-space. A generic (nonsingular) point of the discriminant hypersurface is represented in the physical space by an immersed curve γ with one point of transverse self-intersection. A small displacement of the point from the discriminant hypersurface in a direction transverse to it, transforms the curve γ into an embedded curve γ^+ or γ^-. The self-intersection point is represented on each of these embedded curves by two points 1 and 2. The velocity vectors of the embedding at the points 1 and 2, together with the vector 12, form a frame in the 3-space. Its orientation is positive in one case (γ^+) and negative in the other (the result does not depend on the choice of the points 1 and 2, for instance, nor on their ordering). □

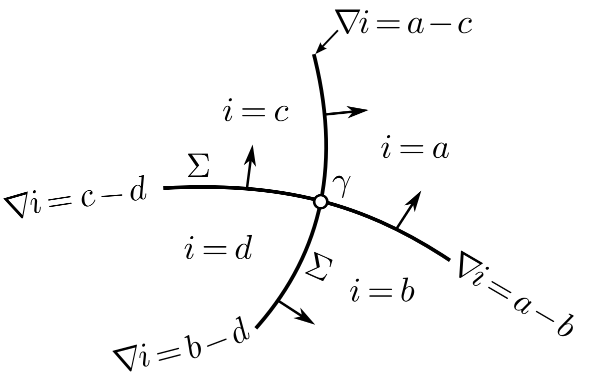
Definition. The *jump* of an invariant i at a point of the discriminant hypersurface is the difference of the values of the invariant, evaluated at both sides of the hypersurface:

$$(\nabla i)\gamma = i(\gamma^+) - i(\gamma^-) \ , \quad i \in H^0(\mathcal{F} - \Sigma) \ .$$

Thus, ∇i is a locally constant function on the set of nonsingular points of the discriminant.

Iterating this construction, one defines the n-th jump, $\nabla^n i$, which is a locally constant function on the set of immersions whose images have n double points.

Example. The second jump of an invariant is defined at the self-intersection points of the discriminant hypersurface as the jump of the first jump of the invariant at the first branch of the discriminant hypersurface – Fig 18.3.



$$\begin{aligned}(\nabla^2 i)\gamma &= (a-b)-(c-d)\\ &= (a-c)-(b-d)\end{aligned}$$

Figure 18.3: The second jump of an invariant is independent of the ordering of the branches of the discriminant variety.

Its value does *not* depend on the choice of the branch of the discriminant hypersurface that was called above the first one. Similarly, the higher jumps are well defined.

Definition. A *Vassiliev invariant of order* n is a knot invariant whose $(n+1)$th jump vanishes identically.

Theorem. *The Vassiliev invariants form a sub-ring of the ring of all knot invariants. Indeed, the following version of the Leibniz formula holds:*

$$\nabla(ij) = i^+j^+ - i^-j^- = i^+j^+ - i^+j^- + i^+j^- - i^-j^- = (i^+)\nabla j + (j^-)\nabla i \ .$$

Hence the product of Vassiliev invariants of orders m and n is a Vassiliev invariant of order at most $m+n$.

Remark. Vassiliev has conjectured that his invariants distinguish any two knots. This conjecture has been neither proved nor disproved. In any case, the Vassiliev invariants distinguish at least as many knots as all other known invariants. For instance, if one substitutes e^t for the variable in the Jones polynomial* and develops the resulting function in a Taylor series, then the coefficient of the term containing t^n will be a Vassiliev invariant of order n (Birman and Lin [50]). Hence all knots distinguished by the Jones polynomials are distinguished also by the Vassiliev invariants. Similar results hold for all other known polynomial invariants.

Remark. The Vassiliev ring has not yet been computed explicitly. However Bar-Natan and Kontsevich announced that the corresponding graded ring (tensored with $\mathbb{C}$) is isomorphic to the graded ring of polynomials in an infinite set of indeterminates whose degrees are such that the number $\#(n)$ of indeterminates of any fixed degree n is finite

n	1	2	3	4	5	6	7	8	9	10	11	12	13
$\#n$	0	1	1	2	3	5	8	12	18	27	39	55	?

One thus finds the dimensions of the spaces of Vassiliev invariants of small order n to be:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$\dim V_n$	1	1	2	3	6	10	19	33	60	104	184	316	548	?

*The Jones polynomial, discovered by V. Jones in 1983, is a knot invariant more sensitive than the Alexander polynomial.

For $n < 5$, these dimensions and spaces were calculated by Vassiliev [132], for $5 \leqslant n \leqslant 9$, they were calculated by Bar-Natan (using many hours of Cray computations) and for $n = 10, 11, 12$, by J.Kneissler [92].

Before we start to calculate the ring of Vassiliev invariants, let us discuss the motivations behind its definition.

The standard technique in the topologic work with discriminant varieties is the following *resolvent* construction. Replace each self-intersection point by two copies of it (one at each branch) and add a segment so that these are joined points. Then replace all the triple points by triads of points and glue a closed 2-simplex to each such triad. Glue 3-simplices to the resolved quadruple points, and so on – Fig. 18.4.

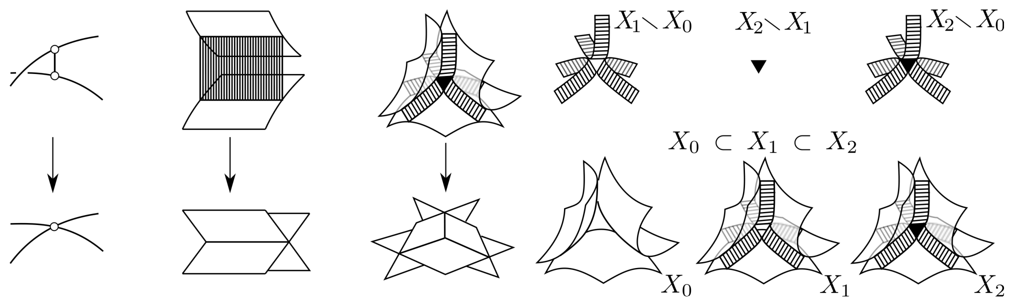


Figure 18.4: The resolution of self-intersections.

The resulting topological space X is homotopy equivalent to the initial one. It has an increasing filtration $X_0 \subset X_1 \subset X_2 \subset \dots$, where $X \setminus X_i$ replaces the self-intersections of multiplicity greater than i. The difference $X_i \setminus X_{i-1}$ is the closure of the space of the fibration into open i-simplices over the set of self-intersection points of multiplicity i. The space X_0 is the closure in X of the set of non self-intersection points of the initial discriminant variety.

Now one considers the spectral sequence associated to this filtration. The Vassiliev iterated jumps occur naturally in the study of the first differential of this spectral sequence (see [132]). If this sequence converges to the cohomology of the space of knots, then the Vassiliev invariants distinguish all knots.

This way of thinking, so natural from the singularity theory point of view, was rather unusual for the knot theorists. Vassiliev theory had not been noticed by the knot theory community until V. Arnold explained it to Joan Birman, and posed the problem of whether Vassiliev invariants distinguish more knots than do the one variable Jones polynomials (a question that she and X. S. Lin subsequently settled affirmatively).

In 1992, Kontsevich stated in his Bonn lectures that the Vassiliev spectral sequence degenerates at the first term (at least when tensored with $\mathbb{C}$).

18.4 Calculation of the Vassiliev invariants

The dual of the free finitely generated abelian group V_n/V_{n-1} admits an explicit combinatorial description: it is generated by the Vassiliev diagrams (defined below), modulo some relations. These relations, described below, are as fundamental as the relations in braid groups, the Jacobi identity and as the Yang-Baxter and Knizhnik-Zamolodchikov equations mentioned above, which are closely related to the combinatorics of the relations between the Vassiliev diagrams. To understand the nature of these relations, we shall start to calculate the Vassiliev invariants of small order n. It is technically convenient to represent the knots by embeddings $\mathbb{R} \to \mathbb{R}^3$ with boundary conditions at infinity (the corresponding functional space of maps $\mathcal{F}$ is an affine space).

Invariants of order 0

The defining relation $\nabla i = 0$ means that the invariant i is constant globally. Hence *the space of zero order Vassiliev invariants is the space of constants*

$$V_0 = \mathbb{Z}$$

(similar to the space of polynomials of degree zero).

Invariants of order 1

The defining equation $\nabla^2 i = 0$ means that *the first jump of the invariant i is constant on all the immersions with just one point of non tangent self-intersection* – Fig 18.5.

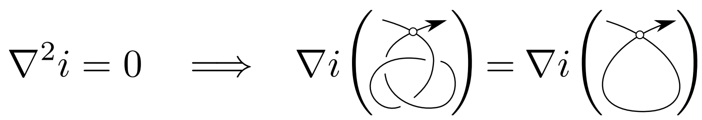


Figure 18.5: The constancy of the jump of an invariant of order 1.

Indeed, each pair of immersions of this class is joined by a finite chain of surgeries ("perestroikas") during which one branch of the curve moves through the other, introducing at that moment one new double point of non tangent self-intersection of the immersed curve. The jump of the first jump at any such surgery vanishes, since $\nabla^2 i = 0$. Hence the value of the jump is the same as for the standard plane curve γ – Fig 18.6.

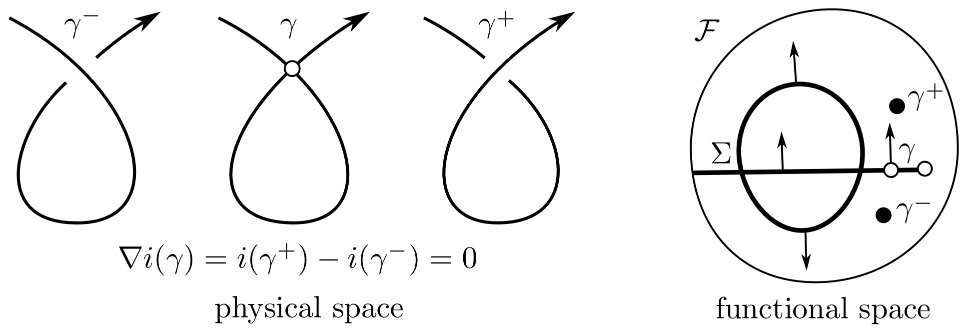


Figure 18.6: Calculation of the jump of an invariant of order 1.

For the standard plane curve γ with one self-intersection point the discriminant is surrounded by the same component of the complement from both sides, since the curves γ^- and γ^+ are regularly isotopic. Hence $(\nabla i)\gamma = 0$, and thus any first order invariant is a zero order invariant:

$$V_1 = V_0 = \mathbb{Z} \; .$$

Remark. In terms of the functional space $\mathcal{F}$, the preceding result expresses the following information on the discriminant hypersurface:

(1) the strata that correspond to more complicated singularities of the discriminant hypersurface, as well as the set of transverse self-intersection of two branches, *do not divide* the discriminant hypersurface.

(2) the stratum of codimension 2 in $\mathcal{F}$, that is, formed by the simplest (cusped) singular curves in $\mathbb{R}^3$, is *the boundary* of the discriminant hypersurface.

The mini-versal deformation of a semi cubic cusp is two-parametric, and the discriminant hypersurface intersects the plane of the parameters along a ray that ends at the point representing the cusped curve. (Figure 18.6 is thus rather realistic).

It is clear that the points of the plane at both sides of a ray belong to the same component of the complement to that ray. This explains the existence of a regular isotopy between the embeddings γ^+ and γ^-, which, moreover, is geometrically evident.

The calculations of the higher order invariants are similar to what we have done; only the simplest information on the stratification of the discriminant hypersurface, corresponding to the hierarchy of singularities, is used. This information is provided by the versal deformations of some few very simple singularities. The next step, where the relevant singularity is the triple point, is crucial for the whole theory.

Invariants of order 2

The defining equation $\nabla^3 i = 0$ means that the second jump of the invariant i does not change when an immersion whose image has two points of non tangent self-intersection undergoes a surgery that introduces for a moment a third self-intersection point.

Unlike the immersions whose image has one self-intersection point, *the immersions with two such points cannot in general be connected by a finite chain of surgeries each of which introduces momentarily one more self-intersection point*.

Indeed consider the preimages of the double points on the oriented line by examining their maps in 3-space. There are 4 preimages and they form two pairs (the two points of a pair have the same images in 3-space).

It is convenient to describe a decomposition of the set $\{1, 2, \ldots, 2n\}$ into n pairs by a system of n arcs lying in the upper half-plane such that the i-th point is connected with the j-th one by one arc if and only if (i, j) is a pair. We shall call any such system of n arcs a *Vassiliev diagram* of order n – Fig 18.7.

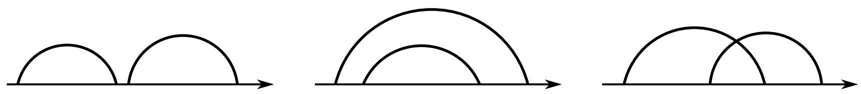


Figure 18.7: The Vassiliev diagrams of order 2.

Of course, many people have previously studied these diagrams, which, for instance, describe the classes of complete flags in a linear symplectic space of dimension $2n$. The components of the knot space are the orbits of the coadjoint representation of $S\mathrm{Diff}\,\mathbb{R}^3$ *, which may be more than just a

*WHAT IS $S\mathrm{Diff}\,\mathbb{R}^3$?

coincidence.

There exist exactly 3 Vassiliev diagrams of order 2 – Fig. 18.7.

The Vassiliev diagram of an immersion with n double points does not change under a surgery that introduces momentarily one more double point of the immersed curve. Hence *there exist at least three immersions of a line with two double points that cannot be reduced to one another by a chain of such surgeries* – Fig. 18.8.

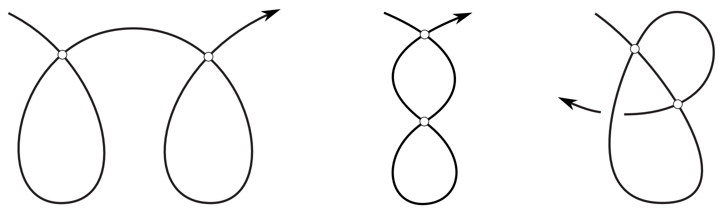


Figure 18.8: The three standard immersed curves with 2 double points.

Any immersed curve with two self-intersection points can be reduced to one of these three standard curves by a finite chain of standard surgeries that introduce momentarily a third self-intersection point.

Therefore, any Vassiliev invariant of order 2 is determined (up to an additive constant) by the values of its second jump on the three standard curves of Figure 18.8.

The values of the second jump of any invariant on the first two standard curves of Fig. 18.8 *vanish.* This follows from the fact that both resolutions of one of the self-intersections produce equivalent (smoothly isotopic) immersed curves with one transverse self-intersection – Fig. 18.9.

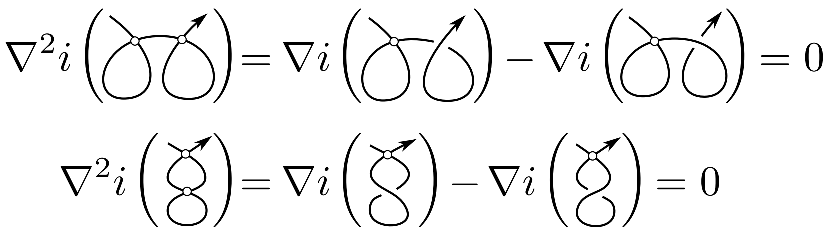


Figure 18.9: Evaluation of the second jump.

Thus, the second jump of an invariant of the second order is unambiguously defined by its value on the third curve. If this value does not vanish (i.e., if the invariant is genuinely of second and not of the first order) then we can multiply it by a constant in such a way that the value on the third curve of Fig. 18.8 will be equal to 1.

A second order Vassiliev invariant v_2 with these properties exists and is unique, up to an additive constant. Thus, $V_2 \approx \mathbb{Z}^2$. We can eliminate the constant by choosing the value of the invariant on the trivial knot (also called *unknot*) to be zero.

Problem. Calculate the value of this invariant v_2 for the trefoil knot.

Solution. The calculation is presented in Fig. 18.10, where K denotes the third curve of Fig. 18.8 (for simplicity, the signs are neglected):

We have assumed that the second jump of v_2 on K equals 1: $(\nabla^2 v_2)K = 1$.

Now, by definition, $(\nabla^2 v_2)K = (\nabla v_2)K^+ - (\nabla v_2)K^-$. Since K^- is equivalent to the γ curve of Fig. 18.6, the curves K^{--} and K^{-+} are regularly isotopic, that is $(\nabla v_2)K^- = 0$. Consequently, $(\nabla v_2)K^+ = 1$.

Again, by definition $(\nabla v_2)K^+ = v_2(K^{++}) - v_2(K^{+-})$, but since the curve K^{+-} is the unknot, the invariant take the value 1 for the trefoil knot K^{++}: $v_2(K^{++}) = 1$.

In a short way:

$$
\begin{aligned}
&1 = (\nabla^2 v_2)(K) = (\nabla v_2)(K^+) - (\nabla v_2)(K^-) \ , \quad (\nabla v_2)(K^-) = 0 \ ; \\
&1 = (\nabla v_2)(K^+) = v_2(K^{++}) - v_2(K^{+-}) \ , \quad v_2(K^{+-}) = 0 \ ; \\
&1 = v_2(K^{++}) = \text{ the value of the invariant evaluated at a trefoil knot.}
\end{aligned}
$$

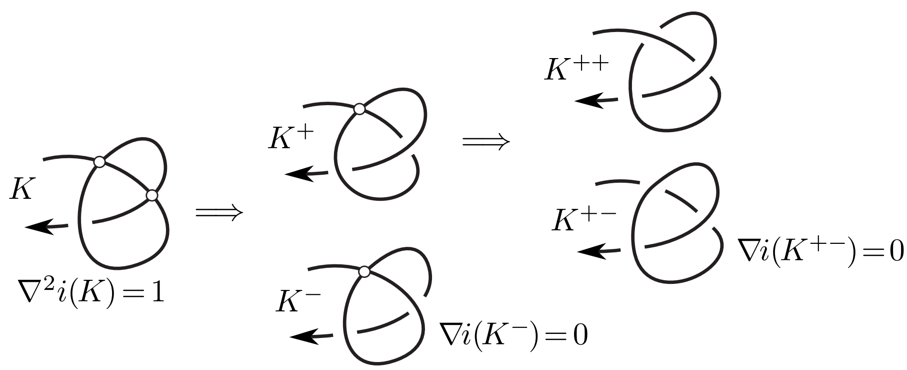


Figure 18.10: Calculation of the Vassiliev invariant of order 2.

The existence of this nontrivial invariant (v_2) was proved in [132]. But this invariant can be reduced to the known ones: It is equal to the x^2 coefficient in the Conway version of the Alexander polynomial. In this case Vassiliev's approach gives an algorithm for calculation of an old invariant. In more complicated cases it generates invariants automatically by standard combinatorial calculations, similar to the preceding ones.

EXERCISE. Prove that the invariant v_2 does not distinguish the trefoil knot from its mirror image (it is well known that they are not equivalent).
Hint: Take the mirror image $\widehat{K}$ of the curve K in Fig. 18.10, and, imitating the above procedure, show that $v_2(\widehat{K}^{--}) = 1$.

18.5 The group of diagrams

The calculation of the Vassiliev invariants of order n is similar to the calculations of those of order 2. The defining equation $\nabla^{n+1} i = 0$ means that the n-th jump of the invariant i is a function that is locally constant on the space of the immersions whose images have n self-intersection points and that does not change under the surgeries that introduce momentarily one more double point. It follows that *the n-th jump, $\nabla^n i$, depends only on the Vassiliev diagram of the immersion with n double points.*

The number of Vassiliev diagrams formed by n arcs is equal to*

$$(2n-1)!! = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n-1) \ .$$

An invariant of order n is defined by the values of its n-th jump on those diagrams up to the addition of an invariant of a smaller order. Hence we obtain the inequality

$$\dim V_n / V_{n-1} \leqslant (2n-1)!!$$

showing that *the space of Vassiliev invariants of any given order is finite dimensional.*

To describe explicitly the space V_n/V_{n-1}, it is convenient to start from the free abelian group $\mathbb{Z}^{(2n-1)!!}$, whose generators are the diagrams of order n.

*$(2n-1)!!$ is the number of the symplectically nonequivalent complete flags $V^0 \subset V^1 \subset V^2 \subset \ldots \subset V^{2n}$ in the symplectic $2n$-space. Have they any unformal relation to the Vassiliev diagrams? Is there a symplectic structure somewhere in knot theory or in the theory of diagrams?

The n-th jump of an invariant of order n is a linear function on the additive group generated by the diagrams. However, as we have seen above for $n = 2$, some of these linear functions are not equal to the n-th jump of any n-th order Vassiliev invariant. For example, for $n = 2$ the values of this function on the first two diagrams of Figure 18.7 must vanish.

In the general case of arbitrary n, the admissible linear functions are those that vanish on some special diagrams or linear combinations of diagrams. We shall describe below these diagrams and combinations. It is convenient to introduce the following definition.

Definition. The *group of diagrams of order n* is the abelian group, denoted by A_n, whose generators are the Vassiliev diagrams consisting of n arcs and whose subgroup of relations (in the free abelian group generated by the diagrams) is generated by the two types of relations described below:

$$A_n = \frac{\mathbb{Z}^{(2n-1)!!}}{(\text{relations 1 and 2})} .$$

Relation 1 (The *easy relations*)**.** Each diagram, containing an arc joining two neighbouring points belongs to the subgroup of relations – Fig. 18.11.

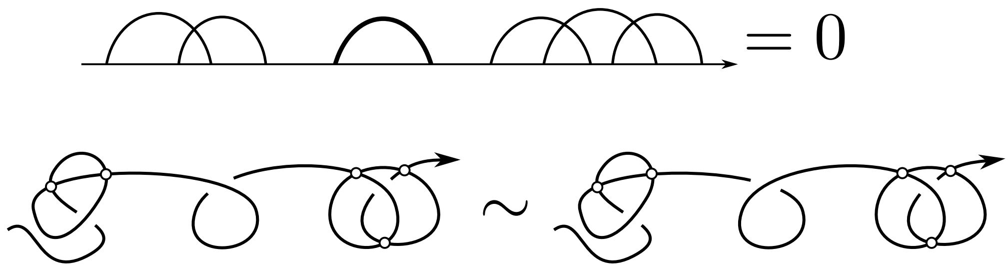


Figure 18.11: An easy relation and its motivation.

Relation 2 (The 4-*term relations*)**.** The combination of four diagrams

$$S_1 - S_2 + S_3 - S_4$$

belongs to the subgroup of relations. The diagrams S_i considered here, consisting of n arcs, are described below (in Figure 18.13).

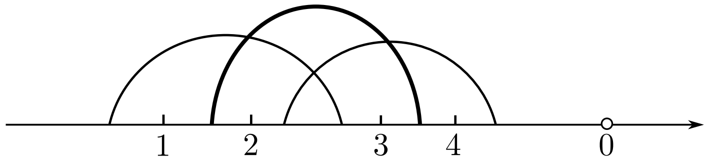


Figure 18.12: Construction of a 4-term relation.

The 4-term relation is a fundamental combinatorial relation whose role in Vassiliev invariants theory is similar to that of the Jacobi identity in Lie algebra theory.

To write the Jacobi identity as a system of relations between the structure constants, we have to fix the value of four indices (i, j, k, ℓ) and then add the corresponding products of the structure constants having those indices. Thus the Jacobi identity is in fact a family of numerical equations, which is parametrised by the choice of indices. The parameter of the 4-term relations of the group A_n consists of the following data:

(1) a Vassiliev diagram of order $n-2$ (shown in Fig. 18.12 for $n = 4$ by the ordinary lines);

(2) one more distinguished arc in the upper half-plane (shown in Fig. 18.12 by a thick line); and

(3) one distinguished point on the border line (0 in Fig. 18.12).

Thus the total number of points at the border line is $2n-1$. These points divide the line into parts. Let us consider the four parts adjacent to the endpoints of the distinguished arc (some of these parts may coincide). We denote them by the numbers $(1, 2, 3, 4)$ in the order defined by the orientation of the border line.

We denote by S_i the diagram obtained as the union of the $n-1$ arcs defined by the above data and of one more arc joining the distinguished point to a point of the part i. A 4-term relation, which corresponds to the data in Figure 18.12, is represented in Figure 18.13.

Remark. The 4-term relations, which were implicit in Vassiliev's initial work [132], have been written in the form described above by Birman and Lin [50].

The number of independent relations among the relations 1 and 2 is at present known only for small n. According to the computations of Vassiliev

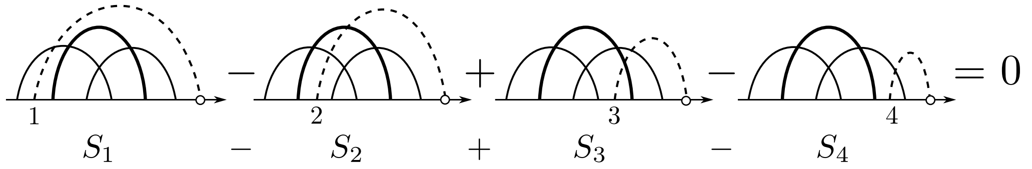


Figure 18.13: A 4-term relation.

($n < 5$), Bar-Natan ($5 \leqslant n \leqslant 9$) and Kneissler, the ranks of the free abelian groups A_n are given by the following table:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$\dim A_n$	0	0	1	1	3	4	9	14	27	44	80	132	232	?

Any function on the set of diagrams with n arcs defines a linear function on the free abelian group generated by the diagrams.

Theorem 1. *The value of the n-th jump of any Vassiliev invariant of order n vanishes on each relation of the diagram group A_n.*

Proof. Fix an easy relation, that is, a diagram containing a short arc. Consider an immersed curve with n double points whose diagram has a short arc. Introducing one more double point at the moment of the surgeries, we can transform this immersion into an immersion for which the short arc is represented by a standard short simple loop in a ball of 3-space containing no other parts of the curve – Fig. 18.14.

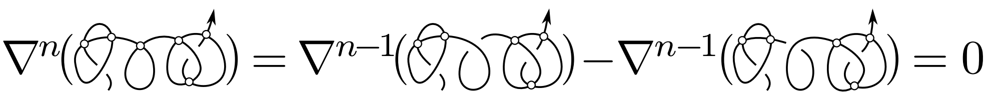


Figure 18.14: Evaluation of the n-th jump on an easy relation.

The values of the n-th jump of any invariant on such a curve is equal to the difference of the values of the preceding jump on two regularly immersed isotopic curves with $n - 1$ double points; hence it vanishes. Theorem 1 is proved for the easy relation.

The 4-term relation appears naturally in the study of the *generic* triple points of immersions (where the tangents of the three branches are three linearly independent lines). Such points occur unavoidably in generic 3-parameter families of maps of a curve in 3-space. The maps with a triple point form a variety of codimension 3 in the space of maps. Its transverse 3-space intersects the discriminant hypersurface along three surfaces intersecting each other transversely – Fig. 18.15. The first surface corresponds to the first return to a point visited by the immersion. The second and third surfaces correspond to the subsequent return to one of the two intersecting branches of the immersed curve (visited at the first and at the second instances).

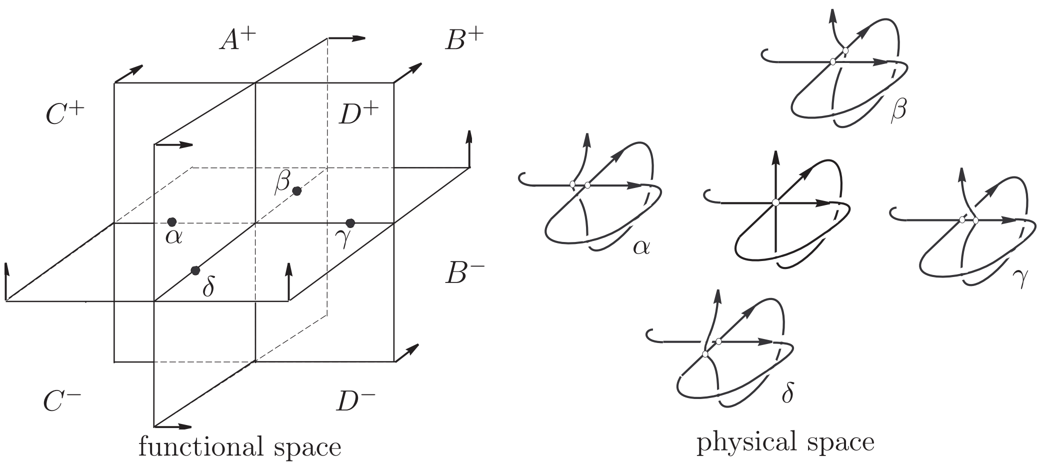


Figure 18.15: The origin of the 4-term relations: deformations of a triple point.

Deform slightly the immersion near the third visit in such a way that the deformed part will intersect one of the 4 rays of the cross formed at the initial self-intersection. The four deformed immersions are shown in Fig. 18.15.

These four deformed immersed curves are represented in the functional space (and in the versal deformation 3-space, shown in Figure 15) by four points $(\alpha, \beta, \gamma, \delta)$ belonging to the codimension 2 strata of the discriminant hypersurface (namely, to its simple self-intersection strata). All these four points belong to one of the branches of the discriminant hypersurface (represented in Fig. 18.15 by a horizontal plane).

Problem. Calculate the second jumps of an invariant i at these points, that is, at the four curves $\alpha, \beta, \gamma, \delta$. Denote their respective jumps by $\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}, \widehat{\delta}$ and, then, prove that

$$\widehat{\alpha} - \widehat{\beta} - \widehat{\gamma} + \widehat{\delta} = 0.$$

Solution. Take the notation $A^{\pm}$, $B^{\pm}$, $C^{\pm}$, $D^{\pm}$ of Fig. 18.15 for the values of the invariant i on the eight octants in the versal deformation 3-space. By definition

$$\begin{aligned}
\widehat{\alpha} &= (A^+ - A^-) - (C^+ - C^-) \ , \\
\widehat{\beta} &= (A^+ - A^-) - (B^+ - B^-) \ , \\
\widehat{\gamma} &= (B^+ - B^-) - (D^+ - D^-) \ , \\
\widehat{\delta} &= (C^+ - C^-) - (D^+ - D^-) \ .
\end{aligned}$$

Evidently, $\widehat{\alpha} - \widehat{\beta} - \widehat{\gamma} + \widehat{\delta} = 0$.

End of the proof of Theorem 1. The relation between the values of the n-th jump of a Vassiliev invariant of order n on the four diagrams S_i (Fig. 18.13) follows from the same arguments, applied to the four deformations of an immersion having one triple point and $n-2$ double points. The value of the n-th jump of an invariant on the deformed immersion with n double points can be considered as the second jump of the $(n-2)$-th jump (as of a locally constant function on the space of maps with $n-2$ double points). Theorem 1 is proved. □

Remark. The parameters of the corresponding 4-term relation have the following meaning. The $n-2$ arcs form the diagram of an immersion in which the triple point disappears completely. The distinguished arc corresponds to the first return to the triple point (preserved under all the four deformations). The distinguished point describes the place of the last return among the moments of the other visits of the double points.

Theorem 1 implies the following important result

Theorem (Vassiliev)**.** *The nth jump of any* (rational) *Vassiliev invariant of order n is a homomorphism $A_n \to \mathbb{Q}$. Moreover, any Vassiliev invariant of order n is defined by this homomorphism, up to the addition of an invariant of smaller order.*

Kontsevich has stated that all the relations between the values of the n-jumps follow from the relations 1 and 2 above:

Theorem (Kontsevich)**.** *Any homomorphism $A_n \to \mathbb{Q}$ is the n-th jump of some Vassiliev invariant of order n.*

In other words,

$$(V_n/V_{n-1}) \otimes \mathbb{Q} \approx \mathrm{Hom}\ (A_n, \mathbb{Q}) \ .$$

Kontsevich's proof, based on complex integration, is sketched in Section 18.6 below. In the original approach of Vassiliev [132], the existence of his invariants was proved by purely combinatorial methods.

Remark (cyclic invariance). The element of the group of diagrams that corresponds to an immersed closed curve with n double points, is well defined: *It does not depend on the place where we cut the circle to obtain the line that we have used in the construction of the diagram.*

Indeed, given any diagram, replace the leftmost point by a new point at the extreme right and connect this new point by a new arc to the right end of the destroyed leftmost arc (see Figure 18.16).

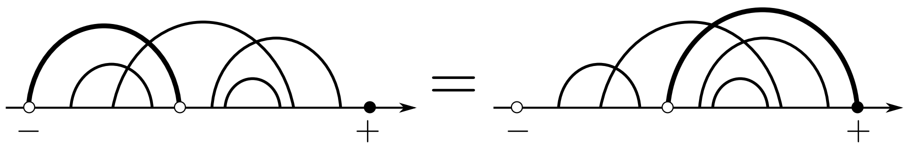


Figure 18.16: The cyclic invariance of a diagram class in A_n.

Lemma. *The new diagram defines the same element of the group of diagrams as the old one.*

Proof. Eliminate from the diagram the leftmost arc. Then, consider each arc, of the $n-1$ remaining ones, as a distinguished arc, and take the 4-term relation associated to it, where the distinguished point is the right end of the eliminated (leftmost) arc. Summing the obtained $n-1$ 4-term relations, one gets the equivalence of the diagrams, since all diagrams appear twice with opposite signs, excepted the two diagrams we are comparing. $\square$

Problem. Prove that *any diagram containing an "isolated arc"* (an arc that does not intersect any other arc) *is equal to zero in the group of diagrams.*

Solution. Using the same reasoning as in the above proof, one can "transport" the left end of this *isolated arc* towards the proximity of its right end "by jumping over the intermediate arcs", coming then to a diagram containing a short arc. To do this, one eliminates from the diagram the isolated arc. Then, in this case, one considers each intermediate arc as a distinguished arc, and takes the 4-term relation associated to it, where the distinguished point is the right end of the eliminated isolated arc. See also Fig. 18.17.

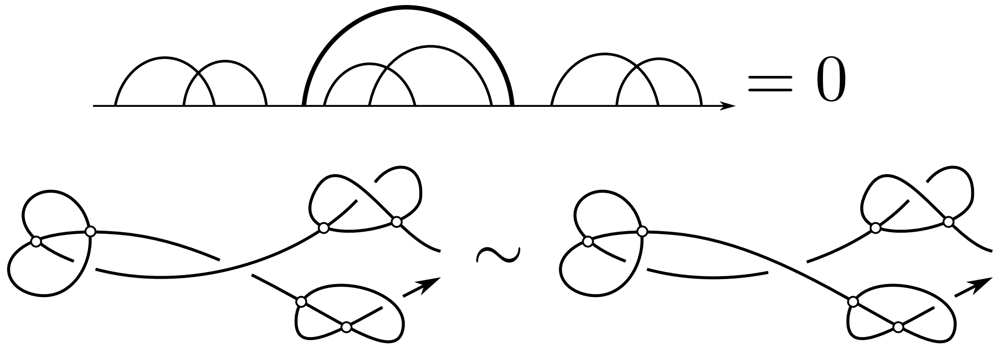


Figure 18.17: A corollary of relations 1 and 2.

Remark. One can combine the diagram groups to form the *diagram ring* $A = \oplus A_n$ by defining the product $A_m \otimes A_n \to A_{m+n}$ as the concatenation of corresponding diagrams – Fig. 18.18.

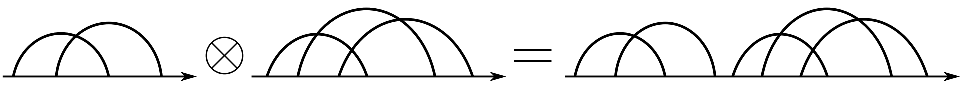
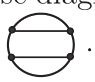


Figure 18.18: Multiplication of diagrams.

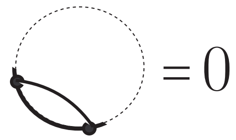
This ring is commutative (by the cyclic invariance proved above). In fact A has also a structure of a commutative and cocommutative Hopf algebra (the comultiplication is dual to the multiplication of the Vassiliev invariants). The elements of the ring A (or rather of its completion) may be viewed as models of knots – the Vassiliev invariants define linear functions on it.

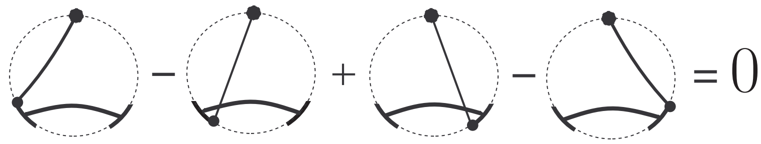
According to a general theorem of algebra, the graded algebra A is isomorphic to the algebra of polynomials. It would be interesting to represent the multiplicative generators by linear combinations of special knots. The arithmetic properties of the coefficients of these combinations are also interesting.

Chord diagrams. By the cyclic invariance (implied by the 4-term relation), the first two diagrams of Fig. 18.7 are equivalent. Indeed, closing the horizontal axes of these diagrams to a circle, both Vassiliev diagrams become the "chord diagram"  .

So, given an immersion $\mathbb{S}^1 \to \mathbb{R}^3$ with n double points, consider the preimages of the double points in the circle, and represent the decomposition of the set $\{1, 2, \ldots, 2n\}$ into n pairs by a system of n chords of the circle such that the ith point is connected to the jth one if and only if (i, j) is a pair. Any such diagram of n chords is called a *chord diagram of order* n^*.

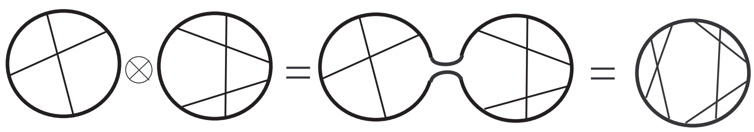
Here, again, the group of chord diagrams of order n is the abelian group, denoted by C_n, whose generators are the chord diagrams of order n, considered modulo the easy relations and the 4-term relations:

Easy relations: 

4**-term relations**: 

Here, we have not drawn all the n chords, and it is supposed that the end points of the undrawn chords lie on the dotted arcs of the circle (that is, out of the small continued arcs). Moreover, for the diagrams in the 4-term relations, it is understood that the $n-2$ undrawn chords are identical in all four diagrams. The 4-term relations are valid for any $n \geqslant 2$, does not matter the position of the supplementary $n-2$ chords.

The direct sum $C = \oplus C_n$ is a graded algebra in which the product of two chord diagrams is defined by their connected sum:



This product is well defined and commutative due to the 4-term relations. In fact C is also a bialgebra, and the 4-term relations imply that the bialgebra C (of chord diagrams on a circle) and the bialgebra A (of Vassiliev diagrams on a line) coincide. So we can use the Vassiliev diagrams or the chord diagrams to our convenience.

Invariants of order 3

Observe that there are exactly five chord diagrams of order 3:

---

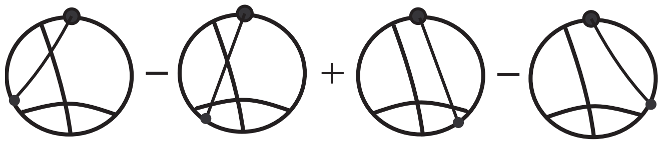
*More formally, a chord diagram of order n is a circle provided with n unordered pairs of points (up to diffeomorphisms of the circle).

.

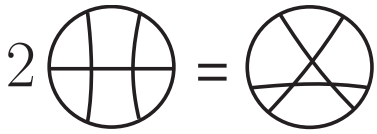
By the easy relations, the first three diagrams are zero in the group C_3. Consequently, the dimension of C_3 is at most 2.

Problem. Prove that the dimension of C_3 (of A_3) equals 1 (see page 688).

Solution. Consider a 4-term relation for $n = 3$ and add a chord to the four diagrams of the above 4-term relation:

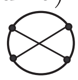
 $= 0.$

Since the last diagram is zero (by the easy relations), we have

, $(*)$

proving that $\dim C_3 = 1$.

So, the 3rd jump of a 3rd order Vassiliev invariant is unambiguously defined by its value on any curve whose chord diagram is the last diagram of equation $(*)$ (for instance, on the curve K of Fig. 18.19). We can fix it to be 2. Such 3rd order Vassiliev invariant exist but is not unique.

We determine uniquely such an invariant v by choosing: a) the value of v on the unknot (for instance, equal to zero); and b) the value of its second jump on a fixed curve whose chord diagram is , for instance, on the curve K^+ of Fig. 18.19.

Remark. The value of the 2nd jump of a 3rd order invariant depends on the choice of the curve having the given diagram, not only on the diagram.

Problem. Prove that the third order invariant v distinguishes the trefoil knot from its mirror image (see the exercise of page 685).

Solution. We consider the successive resolutions of the curve K of Fig. 18.19, for which we have assumed $(\nabla^3 v)K = 2$.

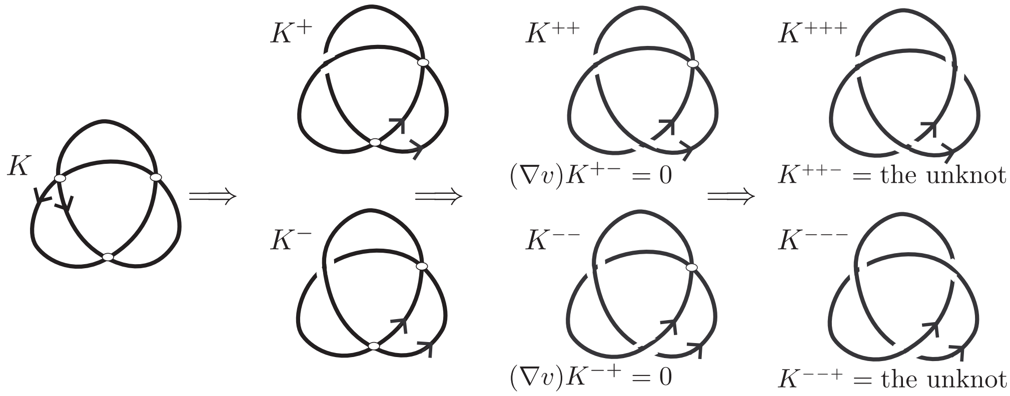


Figure 18.19: The 3rd order invariant v distinguishes the two trefoils.

Draw the curves K^{+-} and K^{-+}, and observe that both resolutions of the curve K^{+-} (and those of K^{-+}) are smoothly isotopic to the unknot, that is, $(\nabla v)K^{+-} = 0$ (and $(\nabla v)K^{-+} = 0$). Similarly, both the curves K^{++-} and K^{--+} are smoothly isotopic to the unknot, that is, $v(K^{++-}) = v(K^{--+})$. Using these three equalities we get

$$\begin{aligned}(\nabla^3 v)K &= (\nabla^2 v)K^{+} - (\nabla^2 v)K^{-} \\ &= [(\nabla v)K^{++} - (\nabla v)K^{+-}] - [(\nabla v)K^{-+} - (\nabla v)K^{--}] \\ &= [v(K^{+++}) - v(K^{++-})] + [v(K^{--+}) - v(K^{---})] \\ &= v(K^{+++}) - v(K^{---})\end{aligned}$$

that is, $v(K^{+++}) \neq v(K^{---})$.

18.6 Kontsevich integrals for Vassiliev Invariants

In the beginning of the 90's M. Kontsevich presented some explicit formulas for the Vassiliev invariants of order n in a form of n-dimensional integrals, similar to the Gauss integral for the linking number discussed in page 386.

To explain them, we consider $\mathbb{R}^3$ as the product of the *horizontal plane* of a complex coordinate z and of the *vertical axis* of a real coordinate t. Moreover, we represent a knot K as a *"nice Morse embedding"* $\mathbb{S}^1 \to \mathbb{R}^3$, for which all the critical points of the restriction of t to the knot curve are Morse non-degenerate and all the critical values are different.

The construction starts from the iterated integrals defining the *Morse knot invariants*, which are constant along the components of the set of embeddings having only Morse critical points.

Choose n non critical values $t_1 < \ldots < t_n$. Choose two different points (z_i, z_i') among the points of intersection of the knot with the horizontal plane $t = t_i$ – Fig. 18.20.

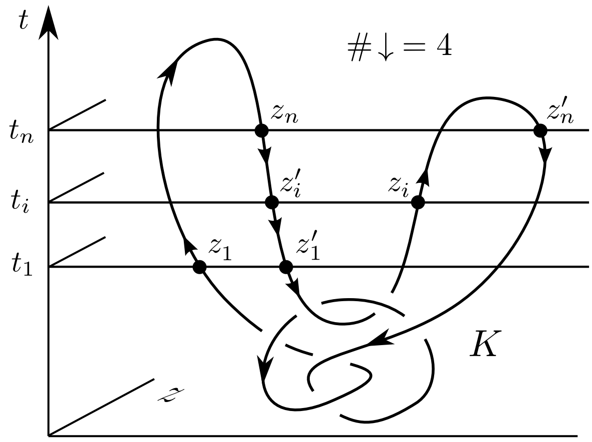


Figure 18.20: The construction of the Kontsevich integral.

The knot branches define locally n pairs of smooth curves $z_i(t), z_i'(t)$ on the real plane of the complex coordinate z. The n-times iterated Kontsevich integral is the integral with values in $A_n \otimes \mathbb{C}$,

$$\widetilde{I}(K) = \int \ldots \int_{t_1 < \ldots < t_n} \sum_{\{z_i, z_i'\}} \left[w \bigwedge_{i=1}^{n} \frac{dz_i - dz_i'}{z_i - z_i'} (-1)^{\#\downarrow} \right] ,$$

where the *weight* $w \in A_n$ represents * the nth order Vassiliev diagram whose n arcs connect the two points of the respective n pairs (z_i, z_i') on the oriented circle K, and where $\# \downarrow$ is the number of *descending points* among $\{(z_i, z_i')\}$ (points where the orientation of K is opposite to that defined by dt). The summation is over all possible choices of the points z_i and z_i' for all i.

Remark. This integral is absolutely convergent. Indeed, the weight w vanishes at a neighbourhood of a zero of the denominator, according to the easy relation 1 in A_n (see Section 18.4).

*THE *WEIGTH* WAS NOT DEFINED AND IT IS NOT CLEAR WHAT MEANS HERE THE WORD “REPRESENTS”

Thus *the integral depends on the Morse embedding K continuously.*

Constancy property. The crucial property of the Kontsevich integral is its *constancy along the deformations of the embedding K in the class of the Morse knots.*

This property depends on the following elementary, but strange, lemma.

Lemma.

$$\frac{dz_1 - dz_2}{z_1 - z_2} \wedge \frac{dz_2 - dz_3}{z_2 - z_3} + \text{cyclic permutations} \equiv 0 \ .$$

Proof. Compute. □

Remark. This identity first appeared in [7] as the generator of the identities in the exterior subalgebra formed by the differential forms on the configuration space $\mathbb{C}^n \setminus \text{diag}$ that are generated (over $\mathbb{C}$) by the standard forms $\omega_{i,j} = d\ell n(z_i - z_j)$. This subalgebra is isomorphic to the cohomology algebra of $\mathbb{C}^n \setminus \cup\text{diag}$.

The above identity is closely related to the Knizhnik-Zamolodchikov equation [91] .

The Kontsevich construction depends on the choice of a closed complex $(n-1)$-form ω on $\mathbb{R}^n_1 \times \mathbb{R}^n_2 \setminus \text{diag}$ that verifies three conditions:

(1) the cohomology class $[\omega]$ is nonzero;

(2) ω is antisymmetric, that is, $\sigma^*\omega = (-1)^n\omega$, where σ is the involution exchanging the factors;

(3) let $\omega_{i,j}$ be the form on $\mathbb{R}^n_1 \times \mathbb{R}^n_2 \times \mathbb{R}^n_3 \setminus \cup(\text{diag})$, induced from the form ω defined on $\mathbb{R}^n_i \times \mathbb{R}^n_j \setminus \text{diag}$ under the natural projection, where $i, j \in \{1, 2, 3\}$; then

$$\omega_{1,2} \wedge \omega_{2,3} + \textit{cyclic permutations} \equiv 0 \ .$$

For $n = 2$ such a form is given by the above lemma:

$$\omega = d\ell n(z - z') \ .$$

For $n > 2$ no smooth form verifying the conditions 1-3 is known. The Kontsevich integrals correspond to a generalised solution in the class of currents. One represents $\mathbb{R}^n$ in the form $\mathbb{C}\times\mathbb{R}^{n-2}$ with coordinates $(z, t_1, \ldots, t_{n-2})$. The solution used by Kontsevich is the current

$$\omega = d\ell n(z - z') \wedge d\vartheta(t_1 - t_1') \wedge \cdots \wedge (t_{n-2} - t'_{n-2}) \ ,$$

where $(\vartheta)(t)$ is equal to 1 for positive t and to zero for negative t.

To prove the deformation invariance of a Kontsevich integral, one writes its variation as an integral of some differential form along K and, then, one uses that *this form vanishes identically,* according to the preceding lemma. We leave the details to the interested reader, pointing out that it is here where the four-term relations will be needed.

The deformation invariance implies the second crucial property of the Kontsevich integral; it can be considered as a "Vassiliev invariant of Morse knots with values in $A_n \otimes \mathbb{C}$". Kontsevich has stated that the n-th *jump of the integral* $\widetilde{I}_n$, *evaluated at a Morse immersion* K with n double points, is equal to *the product of* $(2\pi i)^n$ *with the diagram of this immersion*, considered as an element of the diagram group A_n. The idea is to deform K near the self-intersection points, as shown in Figure 18.21 for $n = 1$. It follows, from the Kontsevich iterated jump formula, that *the* $(n+1)$*-th jump vanishes identically.* Thus $\widetilde{I}_n$ may be considered as a generalised vector-valued Vassiliev invariant of Morse knots.

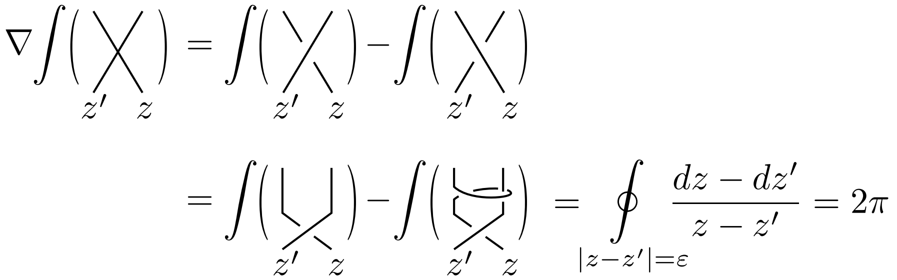


Figure 18.21: The jump as the residue of a Kontsevich integral.

It follows also that *any element of* $A^* \operatorname{Hom}(A_n\mathbb{C})$ *is equal to the* n*-th jump of some complex-valued Vassiliev invariant*:

$$V_n \otimes \mathbb{C} \approx (A_0 \oplus \cdots \oplus A_n)^* \otimes \mathbb{C} \ .$$

To write the Kontsevich formulas for the ordinary knots (which provide invariants independent of the choice of the Morse knot that represents a given knot class), introduce the *total integral* with values in the completion of the algebra $A \otimes \mathbb{C}$,

$$\widetilde{I}(K) = \oplus \widetilde{I}_n(K) \ .$$

Consider an unknotted closed curve K_0 with two Morse maxima and two minima of t. Observe that to calculate the integrals $\widetilde{I}_n(K_0)$, we may replace K_0 by the non closed plane curve $t = x^3 - x + i0$ (Fig. 18.22) since these integrals are invariant under the deformations that do not change the number of maxima of the function t (by the above constancy property).

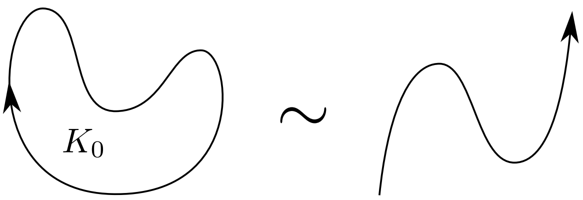


Figure 18.22: The standard curve K_0.

The initial announcement of Kontsevich was wrong because the integrals were non-invariant under some trivial variations of the knot. The error was discovered by Bar-Natan and Kontsevich twisted his integrals to make them unchangeable by those trivial variations. The errors are often more instructive than the proofs!

In order that the Kontsevich integral along an arbitrary Morse knot K be invariant also under the deformations that change the number of maxima of t, Kontsevich had suggested to twist it by using the unknot K_0, in the following way:

$$I(K) := \widetilde{I}(K)/\widetilde{I}(K_0)^m \ ,$$

where m is the number of maxima of t on the knot K.

The division here is understood in the sense of the completion of the algebra $A \otimes \mathbb{C}$: $\widetilde{I}_0(K_0) = 1$ and $(1-a)^{-1} = 1 + a + a^2 + \ldots$, if the order of a in A is positive.

Example. The only number-valued Vassiliev invariant of order 2 (normalised by the conditions that it vanishes on the unknot and takes the value 1 on the trefoil knot) is equal to

$$\Phi(K) = \frac{1}{4\pi^2} \iint\limits_{t_1,t_2} \sum \{z, z'\} \frac{dz_1 - dz_1'}{z_1 - z_1'} \wedge \frac{dz_2 - dz_2'}{z_2 - z_2'} (-1)^{\#\downarrow + \frac{m-1}{6}}$$

where K is a Morse embedding of a circle with $2m$ critical points of t on it and where the summation is over all the choices of the four points (t_i, z_i), (t_i, z_i') $(i = 1, 2)$, such that the points of the first pair $(i = 1)$ alternate with those of the second along the closed curve K.

The Kontsevich integrals equip the ring $V \otimes \mathbb{C}$ with a $\mathbb{Z}^+$-grading, generated by that of A. The splitting of the space V into a direct sum of subspaces isomorphic to V_n/V_{n-1} is done with the help of what is now called "canonical Vassiliev invariants", i.e. invariants obtained by composition of the Kontsevich integral with a weight system of some fixed degree. However the arithmetic properties of this transcendental * grading are not clear.

Conjecturally * the values of $I_n(K)$ belong to $(2\pi i\mathbb{Q})^n \otimes A_n$.

This arithmetics reflects the arithmetical nature of the constants involved in the formulas for the integer-valued invariants (like $4\pi^2$ and $1/6$ in the preceding formula). These constants depend on the values of the D. Zagier *ζ-functions of several variables* at the positive integer points,

$$\zeta(a_1, \ldots, a_n) := \sum k_1^{-a_1} \ldots k_n^{-a_n}$$

(the summation over the integer points in the Weyl chamber $0 < k_1 < k_2 < \cdots < k_n$).

The integer linear combinations of the numbers $\zeta(a)$ form a ring Z. Kontsevich has stated that the values of his integrals on any knot, $\widetilde{I}_n(K)$ and $\widetilde{I}(K)$, belong to $A_n \otimes Z$.

To see how the ζ-function enters in the Kontsevich formulas, it suffices to consider the simplest case of the double Kontsevich integral

$$i_2(K_0) = 1 + \frac{1}{4} + \frac{1}{9} + \cdots = \frac{\pi^2}{6}$$

for the plane curve $K_0 : t = x^3 - x,\ z = x + i0$ – Fig. 18.23.

Below, we consider the points of any pair $\{z_i, z_i'\}$ as being unordered (otherwise one has to multiply the 2^n).

The choices containing a pair $\{z_1, z_1'\} = \{a_2, a_3\}$ or $\{z_2, z_2'\} = \{b_1, b_2\}$ are not admissible, since the corresponding quadruples cannot alternate. The remaining 4 possibilities of the choices provide 4 terms in the integrand, of the form

$$d\ell n a_{1,2} \wedge d\ell n b_{1,3}(-1) + \ldots$$

---

*WHAT TRANSCENDENTAL GRADING MEANS?

*COMPARAR CON LAS REPUESTAS DE DUZHIN

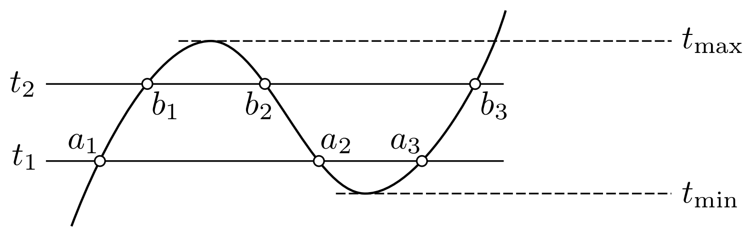


Figure 18.23: The value $\zeta(2) = \pi^2/6$ as a Kontsevich integral.

where $a_{1,2} = a_1 - a_2$, and so on. Taking the signs into account, one can reduce the integrand to the form

$$d\ell n(a_{1,2}/a_{1,3}) \wedge d\ell n(b_{2,3}/b_{1,3}) \ .$$

This expression explains the rather mysterious *invariance of the integral under the deformations of K_0 in the class of Morse embeddings.*

Indeed, the integration domain is the triangle

$$t_{\min} < t_1 < t_2 < t_{\max}$$

and the ratios $u = a_{1,2}/a_{1,3}$, $v = b_{2,3}/b_{1,3}$ send the boundary of this triangle onto the boundary of the standard triangle $u + v \geqslant 1$, $u \leqslant 1$, $v \leqslant 1$, which is invariant under the deformations in the class of Morse embeddings – Fig. 18.24.

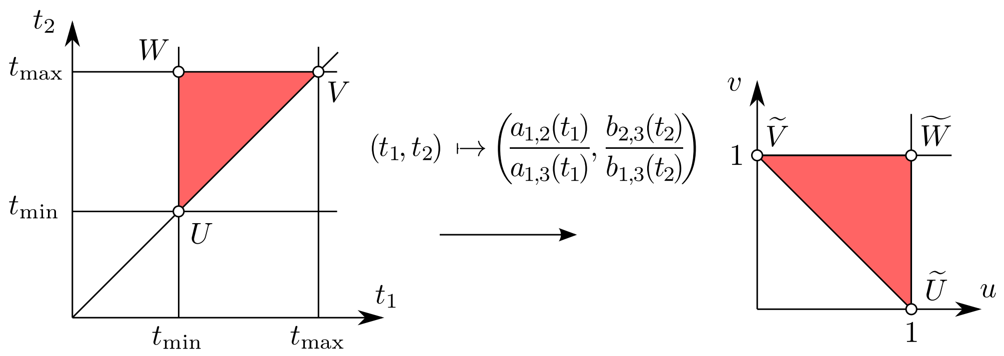


Figure 18.24: Integration domains for the Kontsevich integral equal to $\zeta(2)$.

Thus the integral is reduced to the standard integral along the standard triangle,

$$\iint d\ell nu \wedge d\ell nv = \int_0^1 \ell n(1-u)\frac{du}{u} \ .$$

Now, taking the Taylor expansion of the logarithm, we obtain the value

$$i_2(K_0) = \sum_{k=1}^{\infty} \int_0^1 \frac{u^k}{k} \frac{du}{u} = \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} = \zeta(2) \ .$$

Similar wonderful cancellations are responsible for the independence of other Kontsevich integrals on the choice of the Morse representative in a knot class. The *standard integrals* occurring in these calculations always have the form

$$\iint_{0<t_1<\cdots<t_N<1} \frac{dt_1}{1-t_1} \frac{dt_2}{t_2} \cdots \frac{dt_{a_1}}{t_{a_1}} \frac{dt_{a_1+1}}{1-t_{a_1+1}} \cdots \frac{dt_n}{t_n}$$

(n groups similar to the first product of a_1 forms.) *This number is the value of* $\zeta(a_1, \ldots, a_n)$.

It is clear that the theories described above will be soon developed in many directions.

Vassiliev has started from the stabilisation problem ([8], [14], [16]) of the cohomology rings of the complements to the discriminants and to the caustics in the complex versal deformation spaces of critical points of holomorphic functions of n complex variables ([129], [39], [40], [133]).

Then, applying his methods to the real functions of one variable with restricted singularities (sec [23], [131]) Vassiliev has realised that these also work for the vector functions, for instance, for the knots. Vassiliev has also discussed the applications of his theory to the higher dimensional embeddings.

Bar-Natan, Lin and Kontsevich have defined Feynman diagram groups, starting from more general Feynman diagrams than those of Vassiliev, and using more relations (inspired by the Jacobi identity in Lie algebras). The resulting diagram groups are isomorphic to those of Vassiliev.

Birman, Lin, Bar-Nathan and Kontsevich have used these constructions to associate a Vassiliev invariant to any representation of a simple Lie algebra; Kontsevich has promised applications to the topology of 3- and 4-manifolds, to the cohomology of infinite-dimensional Lie algebras and to associative algebras.

The success of the singularity technique in knot theory should not obscure the fact that many fundamental problems of the topology of the functional spaces of maps with restricted singularities are still open both in the real and in the complex domain, even for the functions of one variable (see [14], [16], [23], [131], [25], [26], [27]).

The singularity theory study of the discriminants and their stratifications has many applications, and Vassiliev's theory of finite knot invariants is only one of them. Lots of examples of other problems leading to the study of complements to discriminants are considered in [134].

Another series of examples is provided by global singularity theory studying topological obstructions to removal of singularities in various contexts. A simplest example is the Poincaré-Hopf theorem relating the Euler characteristic of a manifold to the singular points of a generic vector field on that manifold. In general, global singularity theory studies cohomology classes dual to the loci of singularities of differentiable maps and other kinds of degenerations (caustics, wave fronts). It is established that this cohomology classes are expressed as a universal polynomial (the so called Thom polynomial) in the characteristic classes of the manifolds taking part in the problem. Universal polynomials have been computed explicitly for many particular classes of local singularities (Porteous, Ronga, Damon, Rimanyi, ...) as well as for multisingularities (Kleiman, Kazarian). The global singularity theory covers the problems of both real and complex geometry. Among the recent achievements in the real problems, we would mention the formula of Saeki and Yamomoto [109] that expresses the signature of an oriented 4-manifold as an algebraic number of fibres with certain kind of degeneration for a generic map of this manifold to the 3-space.

Other probable domains of application include symplectic and contact geometry and the theories of immersed plane curves and of evolution of wave fronts.

Even the algebraic geometry of the Plücker formulas, which relate the topology and the singularities of an algebraic curve and of its dual in $\mathbb{C}\mathrm{P}^2$, is still waiting its genuine higher dimensional version, in spite of the 40 relations between the 46 numerical invariants listed in [110] that relate the singularities to some projective differential properties of two dual surfaces in $\mathbb{C}\mathrm{P}^3$.

The singularities of the differential forms dual to the strata of the varieties of different degenerations, deserve a theory that would provide the Guass-Bonnet type formulas for different integer invariants for the natural stratification, including, say, the representatives of the Pontryagin classes as well as the formulas counting higher inflections.

A lot of publications on the Vassiliev theory have appeared, but the list is so long that we only refer to the web site of Bar-Natan.

Chapter 19

Problems Fair

“In mathematics the art of proposing
a question must be held of higher value
than solving it.” GEORG CANTOR

The problems and exercises below, which have different levels of difficulty, has been used for written examinations and form a sample of the kind of problems that we expect the reader should be able to solve. Readers can use them not only to check whether they have mastered the subjects discussed along the book, but also as a source of new information. The process of solving them may provide a more deep insight of the theory to the readers, who could guess and prove generalisations of such results.

19.1 Exercises and Problems by Chapters

Chapter 1 – Geometry on manifolds.

1. Find the diffeomorphic and non-diffeomorphic manifolds in this list :
A. The cotangent bundle space $T^*\mathbb{S}^2$; B. The tangent bundle space $T_*\mathbb{S}^2$.
C. The manifold of unit tangent vectors $T_1\mathbb{S}^2$; D. The real projective space $\mathbb{R}\mathrm{P}^3$;
E. The submanifold of $\mathbb{C}^3$ (with coordinates (x, y, z)) defined by the equations $x^2 + y^2 + x^2 = 1\,,\ |x|^2 + |y|^2 + |z|^2 = 2\,;$ F. The manifold SO(3);
G. The configuration manifold of a satellite with fixed mass centre O;
H. The manifold of oriented (affine) straight lines in $\mathbb{R}^3$;
I. The manifold of non-oriented (affine) straight lines in $\mathbb{R}^3$;
J. The manifold of great circles on $\mathbb{S}^3$.

2. Find the dimensions, the number of connected components and the orientability (or not) of the following manifolds :
A. Orthogonal group $\mathrm{O}(n)$ and Unitary group $\mathrm{U}(n)$;
B. The group $\mathrm{GL}(n, \mathbb{R})$ of real nondegenerate matrices;
C. The group $\mathrm{GL}(n, \mathbb{C})$ of complex nondegenerate matrices;

D. The manifold of symmetric nondegenerate real matrices of order n, $A^t = A$;
E. The manifold of antisymmetric real matrices of order n, $A^t = -A$;
F. The manifold of Hermitian complex nondegenerate matrices of order n, $\overline{A}^t = A$;
G. The manifold of anti-Hermitian complex matrices of order n, $\overline{A}^t = -A$;
H. The manifold formed by all k-dimensional vector subspaces of $\mathbb{R}^n$;
I. The manifold formed by all k-dimensional projective subspaces of $\mathbb{R}\mathrm{P}^n$;
J. The manifold formed by all k-dimensional affine oriented k-planes of $\mathbb{R}^n$;
K. The variety of all ellipses in the affine plane $\mathbb{R}^2$;
L. The manifold of all circles in the affine plane $\mathbb{R}^2$.

3. Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be the map sending (x, y) to $(u = x + y^2, v = -2x - y^2)$.
A. Draw the image of the line $x = 0$ and the preimage of the half-plane $u > 0$;
B. Is the image of any open set open? Is the preimage of any open set open?
C. Does the derivative of f exist at every point?
D. Compute the matrices of the derivative of f at the points $(1, 0)$ and $(1, 1)$;
E. Draw the image of the set $\{(x, y) : |x - 1| \leqslant 0.1; |y - 1| \leqslant 0.1\}$;
F. Draw the image of the set $\{(h_x, h_y) : |h_x| \leqslant 0.1; |h_y| \leqslant 0.1\} \subset T_{(1,1)}\mathbb{R}^2$ under the derivative of the map f at point $(1, 1)$ computed in D;
G. Find the critical points and the critical values of f.

4. (*Normal map*). Consider the ellipse $\mathscr{E} = \{(X, Y) : X^2 + 2Y^2 = 1\}$ in Euclidean plane $\mathbb{R}^2$ with Cartesian coordinates (X, Y). Denote by $f(x, y, z)$ the distance from a point $z \in \mathscr{E}$ to the point of the plane of coordinates (x, y);
A. Find the critical points of the function f on $\mathscr{E}$ (for fixed values of $(x, y) \in \mathbb{R}^2$) and, among them, identify the local minima and the local maxima;
B. Draw the level lines of the function $F : \mathbb{R}^2 \to \mathbb{R}$, defined as

$$F(x, y) = \min_{z \in \mathscr{E}} f(x, y, z) \ ;$$

C. Draw the gradient-field lines of the function F;
D. Find all the points (x, y), at which F is differentiable;
E. Define the normal map $N : \mathscr{E} \times \mathbb{R} \to \mathbb{R}^2$ sending the pair $(z \in \mathscr{E}, t \in \mathbb{R})$ to the point of the plane $\mathbb{R}^2$ at distance t from z along the (interior) normal to the ellipse $\mathscr{E}$ at the point z. Find the points at which the map N is differentiable.
F. Find and draw the critical points and the critical values of the map N.
G. Draw the image of the closed line $\mathscr{E} \times \{t\}$ by the map N for the values

(i) $t = 0.1$, (ii) $t = 0.5$, (ii) $t = 1$, (iv) $t = 1.5$.

5. Let $f = x^2 + y^2 + z^2$, $\quad g = x^3 + y^3 + z^3$ be two functions on $\mathbb{R}^3$.
A. Find the critical points and the critical values of f and g;
B. Find the values c for which the set $\{(x, y, z) : f(x, y, z) \geqslant c\}$ is open in $\mathbb{R}^3$;
C. Find the values c for which the set $\{(x, y, z) : g(x, y, z) \leqslant c\}$ is compact in $\mathbb{R}^3$;

D. Find the values of c such that the surface $g^{-1}(c)$ is smooth;
E. Find the derivative of the map $A : \mathbb{R}^3 \to \mathbb{R}^2$ that sends the point (x, y, z) to $\big(f(x, y, z), g(x, y, z)\big)$, find its critical points and draw its set of critical values;
F. Write h for the restriction of f to the level surface $g = 1$: $h = f\big|_{g=1}$.

(i) Find the critical points of $h = f\big|_{g=1}$ and identify its local minima and local maxima; Are these local minima (maxima) the genuine minimum and maximum of h? Calculate $\inf h$ and $\sup h$.

(ii) Find the values c for which $h^{-1}(c)$ is: (a) a smooth curve; (b) diffeomorphic to the circle $\mathbb{S}^1$.

Chapter 2: Group Invariance

1. Is the rotation group of the cube isomorphic to the symmetric group S_4 (of the permutations of 4 objects)?

2. Find the subgroups of the rotation group of the cube, consisting of 2, 3, 4, 5, 6, 7, 8, 9 elements. For each subgroup H, calculate the number of the conjugate subgroups ($K = gHg^{-1}$) and distinguish the normal (invariant) subgroups.

Chapter 3: Homotopy groups

1. Is the manifold V of all lines in $\mathbb{R}\mathrm{P}^3$ simply connected? ($\pi_1(V) = \{1\}$?).

2. Find the fundamental group of the manifold of the triaxial ellipsoids in the space $\mathbb{R}^3$ (the three axis being different), or at least the number of its elements.

3. Find the following homotopy groups:
A. $\pi_1(\mathrm{SO}(3))$, $\pi_2(\mathrm{SO}(3))$, $\pi_2(\mathbb{T}^2)$, $\pi_k(\mathbb{T}^n)$;
B. $\pi_1(\mathrm{U}(n))$, $\pi_2(\mathrm{U}(n))$, $\pi_2(\mathbb{S}^1 \cup \mathbb{S}^2)$ with $\mathbb{S}^1 \cap \mathbb{S}^2 = 1$ point;
C. $\pi_1(\mathbb{R}\mathrm{P}^n)$, $\pi_1(\mathbb{C}\mathrm{P}^n)$, $\pi_2(\mathbb{R}\mathrm{P}^n)$, $\pi_k(\mathbb{C}\mathrm{P}^2)$ for $k \leqslant 5$;
D. $\pi_1(M)$, M being the manifold of those quadratic forms in Euclidean space $\mathbb{R}^n$, which have no multiple eigenvalues;
E. $\pi_1(N)$, N being the space of degree n complex polynomials of one complex variable, having no multiple roots;
F. $\pi_k(G)$, G being the complement in the complex vector-space $\mathbb{C}^n$ to the union of its diagonal hyperplanes: $z_m \neq z_k$ for any $m \neq k$ in G;
G. $\pi_1(H)$, H being the space of the Lagrangian vector-spaces $\mathbb{R}^n$ in the standard symplectic space $\mathbb{R}^{2n}$, and $\pi_2(H)$, $\pi_3(H)$ for the case $n = 2$;
H. $\pi_1(P)$, P being the manifold of the non-oriented affine straight lines in $\mathbb{R}^3$;
I. $\pi_1(Q)$, Q being the complement of two general position complex straight lines in $\mathbb{C}^2$ (of n generic complex hypersurfaces in $\mathbb{C}^n$);
J. π_k (for $k \leqslant 3$) of the manifold of complex affine lines of the complex plane $\mathbb{C}^2$.
4. Compute the fundamental group of the manifold of all nondegenerate quadratic forms in $\mathbb{R}^3$, and also in $\mathbb{C}^2$.

5. Let N be the closure in $\mathbb{C}\mathrm{P}^2$ of the subset $y^2 = x^2 + x^3$ of $\mathbb{C}^2 = \{(x,y)\}$. Find the homotopy groups $\pi_1(N)$ and $\pi_2(N)$.

6. Are the maps $\mathbb{S}^3 \to \mathbb{S}^2\colon\ (z,w) \mapsto (z^2/w^3)\,,\ \ (z,w) \mapsto (w/z)$ homotopic? Here, $\mathbb{S}^3 = \{(z,w) : |z|^2 + |w|^2 = 1\} \subset \mathbb{C}^2$, $\mathbb{S}^2 = \mathbb{C}\mathrm{P}^1 = (\mathbb{C}^2 \smallsetminus 0)/(\mathbb{C} \smallsetminus 0)$.

7. Calculate the homotopy group π_2 of the complement to the union of a point and a straight line not containing this point in Euclidean 3-space.

Chapter 5: Geometry of Fundamental Groups.

1 (*Circle*). Let t be the slope of the line $y = t(1+x)$, containing the point$(-1,0)$ of the circle $x^2 + y^2 = 1$, in Euclidean plane $\mathbb{R}^2$ with Cartesian coordinates (x,y);
A. Compute x and y for the other intersection point of the line with the circle;
B. Find all the points on the circle having rational coordinates (x,y);
C. Prove that the genus of the Riemann surface of the circle equals 0;
D. Find all the integer solutions of the Diophantine equation $x^2 + y^2 = z^2$ (generalising the Egyptian triangles $(3,4,5)$ and $(12,13,5)$).

2 (*Rational curves*). Let a curve Γ of the plane be defined parametrically by two rational functions, $\{x = p(t), y = q(t)\}$. Such a curve is called *rational*.
A. Prove that the circle, the ellipses and the hyperbolas are rational curves.
B. Find the rational curves among the constant energy curves of a cubical potential, $y^2 + x^3 - x = E$, i.e., for which values of E are these curves rational?
C. Extend the previous theory of real rational curves in Euclidean plane to the complex and quaternionic cases of "curves" in Hermitian and hyper-Hermitian complex and quaternionic planes, replacing the real circle $\mathbb{S}^1 \approx \mathbb{R}\mathrm{P}^1$ by its complex version $\mathbb{S}^2 \approx \mathbb{C}\mathrm{P}^1$ and by its quaternionic version $\mathbb{S}^4 \approx \mathbb{H}\mathrm{P}^1$. Apply these formulas to the topological study of the groups Spin 3 $\approx$ SU(2) and Spin 5;
D. Prove that any abelian integral along a (real or complex) rational curve is an elementary function (like in the Newton standard integrals cases

$$I(X) = \int_{X_0}^{X} \frac{dx}{\sqrt{x^2 + ax + b}}\,,$$

which would be generically a non-elementary elliptic function if the quadratic polynomial were replaced by a cubical one).

3. Study the 4 complements in $\mathbb{C}^2$ of the following 4 curves: $A : xy = 0$; $B : x^2 = y^4$; $C : x^2 = y^3$; $D : x^2 = y^5$. (a) Which of these complex curves are mutually homeomorphic, and which pairs of the complements are homeomorphic. (b) Compute the fundamental groups of these 4 complements. Which of these fundamental groups are a) finite, b) commutative, c) mutually isomorphic?

Chapter 6: Integration – Differential forms.

1. Find the closed and the exact differential forms among the following forms:
A. $r^n(x\,dy - y\,dx)$ in $\mathbb{R}^2 \smallsetminus 0$ (where $r^2 = x^2 + y^2$);
B. $r^m(x\,dy \wedge dz + y\,dz \wedge dx + z\,dx \wedge dy)$ in $\mathbb{R}^3 \smallsetminus 0$ (where $r^2 = x^2 + y^2 + z^2$);
C. $z^k\,dz$ in $(\mathbb{C} \smallsetminus 0) = (\{z \neq 0\})$; D. $z^a \overline{z}^b\,dz$ in $(\mathbb{C} \smallsetminus 0)$;
E. $\dfrac{dz}{(z-1)^2(z^2-a^2)}$ in $(\mathbb{C} \smallsetminus \{1, a, -a\})$; F. dz/w on the complex curve defined by the equation $w^2 = z - z^3$ in $\mathbb{C}^2$ and in $\mathbb{CP}^2$ (where (z, w) are affine coordinates).

2. Calculate the integrals of the form (b) with $m = -3$ along the spheres $(x-1)^2 + (y-2)^2 + (z-2)^2 = R^2$ for all values of the parameter $R > 0$.

3. Compute the mean values of the function $\ln r$ (where $r^2 = x^2 + y^2$) along all circles $(x-a)^2 + (y-b)^2 = R^2$ of Euclidean plane.

Chapter 7: Lie derivative, Lie algebras and Lie groups.

1. Calculate the order of the differential operator $L_u L_v - L_v L_u$ (where u and v are smooth vector-fields).

2. Calculate, for 3 matrices A, B, C of order $n \times n$, the matrix

$$[[A, B], C] + [[B, C], A] + [[C, A], B] \ ,$$

where $[X, Y] = XY - YX$.

3. Compute a matrix A verifying the equation $e^A = e^B e^C$, in terms of the matrices B, C, $[B, C] = BC - CB$, $[B, [B, C]]$, $[C, [B, C]]$ and of the other iterated commutators of B and C (at least mod $O(|B|^r, |C|^r)$ for $r = 1, 2, 3$).

4. Denote by $[a, b]$ the vector product in an oriented Euclidean 3-space $\mathbb{R}^3$.
A. Calculate the vector $[[a, b], c] + [[b, c], a] + [[c, a], b]$;
B. Deduce (from A) the concurrency of the three altitudes of a triangle in Euclidean plane. Hint: Consider the altitude planes for the trihedral pyramid defined by the edges (a, b, c) in Euclidean space $\mathbb{R}^3$;
C. Associate to the vector a the velocity field v of the points of a rigid body, rotating around O with angular velocity a. Compare $v(x)$ to $[a, x]$;
D. Calculate the differential operator of problem 1 for the vector-fields u and v of velocities a (for u) and b (for v);
E. Deduce from items C and D the relation of the Lie algebra of the group SO(3) to the vector products.

4. Prove that the commutator of two Hamilton vector-fields (on a symplectic manifold) is also a Hamilton vector-field and find its Hamilton function.

5. In a Lie group, compute the Poisson bracket of :
A. Two left-invariant vector fields; B. Two right-invariant vector fields;
C. A left-invariant field and a right-invariant field.

6. Prove that the divergence of the Poisson bracket of two divergence free vector-fields (on a manifold with fixed volume element form) vanishes too.

7. Compare the vector-fields rot $([a,b])$ and $\{a,b\}$ for two divergence free vector-fields in the oriented Euclidean space $\mathbb{R}^3$ ($[a,b]$ being the vector product).

8. For the following 6 groups solve the questions A to G below:
(1) The projective transformations $\mathbb{R}\mathrm{P}^1 \to \mathbb{R}\mathrm{P}^1$;
(2) The linear symplectic maps $\mathbb{R}^2 \to \mathbb{R}^2$;
(3) The volume preserving linear maps $\mathbb{R}^3 \to \mathbb{R}^3$;
(4) The linear transformations $\mathbb{R}^3 \to \mathbb{R}^3$ preserving form $x^2+y^2-z^2$;
(5) The conformal diffeomorphisms of the disc $|z|<1$ of $\mathbb{C}$ onto itself;
(6) The conformal diffeomorphisms of the half-plane $\operatorname{Im} z > 0$ onto itself.

A. Calculate their dimensions and identify which of them are Lie groups;
B. Calculate the Lie algebras of the Lie groups of problem A;
C. Which of the Lie algebras of question B are mutually isomorphic?
D. Find isomorphic groups among the above 6 groups and among their connected components of unity (comparing thus 12 groups);
E. Compare the Lie algebras of question B with the Lie algebra of the Poisson brackets of quadratic forms on the symplectic plane $\mathbb{R}^2$: are they isomorphic to it?
F. Find the adjoint and coadjoint representation orbits of the 6 groups, acting on their Lie algebras and on the dual spaces of these algebras (that is, on 12 spaces);
G. Compute the invariant symplectic structures for the coadjoint representation orbits of item F.

9. Find the eigenvectors and eigenvalues of the linear operator $La = [\lambda, a]$, where λ and a are $n \times n$ matrices, in the following cases:
A. λ is a diagonal matrix with elements λ_k, $a \in \mathbb{C}^{n^2}$ an arbitrary complex matrix;
B. λ and a are real antisymmetric matrices;
C. λ and a are anti-Hermitian matrices (a complex $n \times n$ matrix is called *anti-Hermitian* if its Hermitian conjugate matrix is opposite to it, $A^* = -A$).

10. Compute the number of connected components of the complement to the union of hyperplanes in Euclidean n-space, $\mathbb{R}^n \setminus \bigcup_k H_k^{n-1}$, for:
A. The C_{n+1}^2 hyperplanes of the subspace $\mathbb{R}^n : x_0 + x_1 + \cdots + x_n = 0$ of $\mathbb{R}^{n+1}$, defined by the equations $x_i = x_j$ $(0 \leqslant i < j \leqslant n)$. Draw the pictures for $n=2$ and for $n=3$;
B. The hyperplanes $x_i = 0$ and $x_i = x_j$, $x_i = -x_j$ in $\mathbb{R}^n$ with Cartesian coordinates $x_1, \ldots, x_n$ (the number of hyperplanes is $n + n(n-1)$);

11 (*Weyl chambers*). Compute the number of connected components (Weyl chambers) of the complement in $\mathbb{R}^3$ to the union of the symmetry planes of
A. A regular tetrahedron (the 6 symmetry planes and the Weyl chambers have to be shown in a drawing); B. A cube; C. The octahedron;
D. The icosahedron (which has 30 symmetry planes); E. The dodecahedron.

12. (*Springer numbers*). The three planes that bound a Weyl chamber (of the

symmetry group of a regular polyhedron) decompose the space $\mathbb{R}^3$ into 8 parts. Each of these parts is subdivided by the remaining symmetry planes of the polyhedron. Count the number of parts of these subdivisions for
A. A regular tetrahedron; B. A cube; C. The icosahedron.
Hint: To count these parts, draw the picture representing them in the projective plane $\mathbb{RP}^2$ of the directions in $\mathbb{R}^3$: The symmetry planes become lines and the chambers, triangles. In this projective plane, one identifies the opposite chambers; hence the number of triangles is one half of the number of Weyl chambers.

13. (*Spin*). Consider the action of the group $\mathbb{S}^3 \times \mathbb{S}^3$ (of pairs of quaternions of norm 1) on $\mathbb{H} = \mathbb{R}^4$, defined by the formula $(a, b)(h) = ahb^{-1} \in \mathbb{H}$ and consider the resulting map $A : \mathbb{S}^3 \times \mathbb{S}^3 \to \mathrm{O}(4)$.
A. Find the image of the map A;
B. Find the kernel of the homomorphism A;
C. Study the universal covering map

$$\mathrm{Spin}(4) \longrightarrow \mathrm{SO}(4)$$

(items A and B provide the topological description of the manifold Spin(4)).

Chapter 8: Lobachevsky.

1. Find the infimum of the sum of the interior angles of a triangle among all triangles in Lobachevsky plane.

2. Calculate the Lobachevsky geodesic curvatures of the Euclidean straight lines of the Poincaré model Euclidean half-plane.

3. Is it true that:
A. The three medians of any Lobachevsky plane triangle have a common point?
B. The three altitudes of any Lobachevsky plane triangle have a common point? [A similar question is also interesting for the relativistic de Sitter plane geometry].

4. Find the area of the (infinite) triangle of Lobachevsky plane, whose vertices are three points on the absolute (that is, 3 infinitely far points). Are all these infinite triangles isometric?

5. Sketch the rays emanating from the origin in various directions in a plane where the velocity light is $v(y) = 1/(y^4 - y^2 + 1)$.

The behaviour of such rays is similar (models?) the phenomenon of mirage (explained in Ch. 8) in which the rays present an oscillation near the layer at which the velocity of light is minimal.

6. Sketch the geodesics on the torus, using Clairaut's theorem: the product of the distance to the axis of rotation and the sine of the angle a geodesic makes with a meridian is constant along each geodesic of a surface of revolution.

Chapter 9: Riemannian Geometry.

1. Consider the hyperbolic domain of a generic smooth surface in $\mathbb{R}^3$.
A. Prove that a curve on the surface is tangent to an asymptotic line at a point if and only if its osculating plane is tangent to the surface at that point.
Hint: Write the curve as $t \mapsto \big(x(t), y(t), f(x(t), y(t))\big)$ and use equation (1) of p. 318.
B. Prove **Beltrami-Ennepper Theorem**: *The torsions of the asymptotic curves at a hyperbolic point p of a smooth surface in Euclidean space $\mathbb{R}^3$ is given by the formula $\tau = \pm\sqrt{-K}$*, where K is the Gaussian curvature at p.
Hint: Apply item A and use again equation (1) of p. 318.

Corollary. *The asymptotic curves have no flattening point (torsion zero point) in the hyperbolic domain.*

2. Find the sum of the angles of a triangle of area S:
A. On a sphere of radius R; B. On Lobachevsky plane.
C. Generalise the results of A and B to arbitrary Riemannian surfaces.

3. Draw the geodesics of $ds^2 = a\ dx^2 + dy^2$, with $a = 1 - 3y + 9y^3$, issuing from the origin $x = 0 = y$ in the three following directions: $dy/dx = 1, 2, 3$. Are there conjugated points to $(0, 0)$ along these geodesics? (The three geodesic curves have to be indefinitely extended, and the difference between the three asymptotical behaviours has to be described).

4. Find the Gaussian curvature of the surfaces whose Riemannian metrics are listed below:
A. $g = dx^2 + \sin^2 x\, dy^2$;
B. $g = v(x, y)\, dx^2 + w(x, y)\, dy^2$ with $v > 0$ and $w > 0$.

5. Calculate the Riemannian curvature tensor for
A. The standard metric of the complex projective plane $\mathbb{C}\mathrm{P}^2$;
B. The manifolds SO(3) and SO(4) (equipped with their bi-invariant metrics).

Chapter 10: Degree, Index and Linking

1. Calculate the degrees of the following maps.
A. $\mathbb{C}P^1 \to \mathbb{C}P^1$: $w = p(z)/q(z)$, $\deg(p) = m$, $\deg(q) = n$;
B. $\mathbb{R}^{2n} \to \mathbb{R}^{2n}$: $(x_1, \ldots, x_n; y_1, \ldots, y_n) \mapsto (x_1, y_1, \ldots, x_n, y_n)$;
C. $\mathbb{R}^2 \to \mathbb{R}^2$: $(x, y) \mapsto (x^5 + x^2 y + 1, y^3 + x - 2)$;
D. $f : V \to \mathbb{S}^2$, where V is the submanifold $V = \{(x, y, z) \in \mathbb{R}^3 : F(x, y, z) = 10\}$ with $F = x^2 + 2y^2 - 3z^2 + \sin(x + y + z)$ and $f = (\mathrm{grad}\ F/|\mathrm{grad}\ F|)\big|_{V^2}$.
E. $g : T^2 \to \mathbb{S}^2$: $g(\alpha, \beta) = \dfrac{a(\alpha) - b(\beta)}{|a(\alpha) - b(\beta)|}$, where $a : \mathbb{S}^1 \to \mathbb{R}^3$ and $b : \mathbb{S}^1 \to \mathbb{R}^3$ are two immersions of closed curves with no common points $(a(\alpha) \neq b(\beta))$. Consider, for instance, the case $a(\alpha) = (\cos\alpha\, , \sin\alpha\, , \sin(2\alpha))$, $b(\beta) = (0\, , \sin\beta + 1\, , \cos\beta)$ with $\alpha, \beta \in \mathbb{R} \pmod{2\pi}$.

2. (*i*) Compute the double integral

$$I = \int_0^{2\pi} \int_0^{2\pi} \frac{\big[(da/ds), (db/dt), (a(s) - b(t))\big]}{|a(s) - b(t)|^3}\, ds\, dt$$

for the two closed curves a and b, immersed in Euclidean space $\mathbb{R}^3$, parametrised by $a(s) = (\cos s,\, \sin s,\, 0)$, $b(s) = \big(1 + \sin t,\, 0,\, e^{\sin t}\cos(2t)\big)$. The product $[u, v, w]$ denotes the determinant of the three vectors.

(*ii*) Do disjoint embeddings $a, b : \mathbb{S}^1 \to \mathbb{R}^3$ exist, such that $I = 1$?

3 (*Total Gaussian Curvature, Index of a Vector Field*). In Euclidean space $\mathbb{R}^3$, with Cartesian coordinates x, y, z, compute the integral of the Gauss curvature along the surfaces given by the following equations :
A. $xyz = 1$; B. $z = x^2 + y^4 + \sin x$; C. $z = u(x, y)$, where $u = x^3 - 3xy$.
D. Can the direction field of $\operatorname{grad} u$ be extended by continuity from the affine plane $\mathbb{R}^2$ to the projective plane $\mathbb{R}\mathrm{P}^2$? Is it smooth in $\mathbb{R}\mathrm{P}^2$? Does it have singular points ?
E. Find the index of the vector-field $\operatorname{grad} u$ along the curve $x^2 + y^2 = 1$.

4. Find the index of the vector-field $\dfrac{z}{2 + z^2}$ along the circle $|z| = c$ in $\mathbb{C}$ (for $c = 1$ and $c = 2$).

5. Find the intersection indices for the following pairs of submanifolds.
A. In $\mathbb{R}^2$, the curves $(x = \cos s, y = \sin s)$ and $(x = \sin 3t, y = \cos 2t)$, for $s \in \mathbb{R} \pmod{2\pi}$, $t \in \mathbb{R} \pmod{2\pi}$;
B. In $\mathbb{R}^2$, the curves $x^2 + y^2 = 2 + \sin(xy)$ and $y = 2x$;
C. In $\mathbb{C}^2 = \{(z, w)\}$, two surfaces $\{w = 0\}$ and $\{w = z^2\}$;
D. The surfaces $\{x^2 + y^2 + z^2 = 1\}$ and $\{x^2 + y^2 + z^2 = 1\}$ of $\mathbb{R}^3$ in the manifold (of real dimension 4) of the complex solutions of the same equation;
E. The 2 real compact surfaces defined by the equations $w = 0$ and $zw = 1$ in the real 4-manifold $\mathbb{C}P^2$ (equipped with the affine coordinates (z, w));
F. The two surfaces, defined by the equations $z^2 + w^2 = 1$ and $z^3 + w^3 = 2$ (in the notations of problem E);
G. The two curves defined by the equations $\{x^2 + y^2 = 1\}$ and $\{2x^2 + 3y^2 = 4\}$ in (*i*) $\mathbb{R}^2$, (*ii*) $\mathbb{C}^2$, (*iii*) $\mathbb{R}\mathrm{P}^2$, (*iv*) $\mathbb{C}\mathrm{P}^2$ ((x, y) being affine coordinates in these real and complex affine and projective spaces).

6. Let $M^2 \subset \mathbb{C}\mathrm{P}^2$ be the complex projective elliptic curve, defined in affine coordinates by the equation $y^2 = x^3 - x$. Find the minimal number of intersections of the smooth surface M^2 with a transverse close surface into $\mathbb{C}\mathrm{P}^2$. What happens without the transversality assumption ?

Chapters 10 – 11 (Euler Characteristic)

1. Calculate the Euler characteristic of the following manifolds:
A. The sphere $\mathbb{S}^n$ and the projective spaces $\mathbb{R}\mathrm{P}^n$ and $\mathbb{C}\mathrm{P}^n$;
B. The torus $T^n = (\mathbb{S}^1)^n$ and the Möbius band;
C. The manifold of non-oriented straight lines in the space:
(*i*) $\mathbb{R}^2$, (*ii*) $\mathbb{R}^3$, (*iii*) $\mathbb{R}P^2$, (*iv*) $\mathbb{R}\mathrm{P}^3$ (affine lines for (*i*) and (*ii*));
D. $\mathrm{SO}(n)$, $\mathrm{O}(n)$, $\mathrm{U}(n)$ and $\mathrm{SU}(n)$;
E. The manifolds $\mathrm{GL}(n,\mathbb{R})$ and $\mathrm{GL}(n,\mathbb{C})$ of nondegenerate real and complex $n \times n$ matrices;
F. the manifolds of real symmetric nondegenerate matrices of order 2 and 3;
G. The manifold defined in $\mathbb{C}^3$ by the two equations

$$x^3 + y^4 + z^5 = 0 \ , \quad |x|^2 + |y|^2 + |z|^2 = 1 \, ;$$

H. the complement to n different points in $\mathbb{R}^2$;
I. the sphere $\mathbb{S}^2$ with g handles (“surface of genus g”).

Chapter 11: Homology.

1. Calculate the Betti number b_2 of the following seven manifolds:
A. Complex projective space $\mathbb{C}\mathrm{P}^n$; B. Orthogonal matrix group $\mathrm{O}(3)$;
C. Unitary matrix group $\mathrm{U}(2)$;
D. The hypersurface $\{x^4 + y^4 + z^4 = 1\}$ in $\mathbb{C}^3$;
E. The “Complex sphere” in $\mathbb{C}^3$ defined by the equation $x^2 + y^2 + z^2 = 1$;
F. The manifold $G_{4,2}$ of the vector subspaces $\mathbb{R}^2$ in $\mathbb{R}^4$;
G. The manifold $\widehat{G}_{4,2}$ of the oriented vector subspaces $\mathbb{R}^2$ in $\mathbb{R}^4$.

2. Calculate the signature of the intersection forms for 2-cycles on the manifolds of problem 2 A, 2 C, 2 D, 2 E, 2 G.

3. Compute the Betti numbers of the complement to the union of three complex lines of $\mathbb{C}\mathrm{P}^2$, for all the possible configurations of these lines.

4. Prove the closeness of the differential form dx/y on the submanifold of $\mathbb{C}^2$ $\{y^2 = x(x-1)(x+3)\}$ (with coordinates x, y). Then, compare the integrals of this form along the real arcs $\{-3 \leqslant x \leqslant -1, y \geqslant 0\}$ and $\{0 \leqslant x \leqslant 1, y \geqslant 0\}$: which of these two integrals is larger? (In other terms: is the period of the oscillations in the deeper well larger or smaller than in the less deep one, for the same total energy in a Newton system, whose potential energy is a quartic polynomial?)

Chapter 13: Abel theorem.

1. Calculate the monodromy groups of the following complex algebraic functions $y(x)$: A. $y^2 = x^4 + 1$; B. $y^2 = x^3 + x$;
C. $y^5 + xy + 1 = 0$; D. $y^n + xy + 1 = 0$; E. $y = \sqrt{1 + \sqrt{x}}$.

2. A group G is *solvable* if there exists such a chain of subgroups

$$(G = G_0) \supset G_1 \supset G_2 \supset \cdots \supset G_n = \{1\} \ ,$$

that G_{k+1} is a normal (invariant) subgroup in G_k, the quotient group G_k/G_{k+1} being commutative. Are the following groups solvable?
A. Symmetry group of an equilateral triangle;
B. Symmetry group of a regular tetrahedron (in Euclidean space $\mathbb{R}^3$);
C. The group of the rotations preserving a cube (in Euclidean space $\mathbb{R}^3$);
D. The symmetry group of a cube (in Euclidean space $\mathbb{R}^3$);
E. The group of rotations preserving a regular dodecahedron (in $\mathbb{R}^3$);
F. Symmetry group of a regular icosahedron (in Euclidean space $\mathbb{R}^3$);
G. The symmetric group of the permutations of n elements;
H. The group of even permutations of n elements; I. Group G of the exact sequence $X \to G \to Y$ between two solvable groups X and Y.

3. Consider three complex (multivalued) algebraic functions f, g, h: $y = f(x)$, $z = g(y)$, $z = (g(f(x)) = h(x)$. If f has m values and g has n values, we consider h as having mn values.
A. Let γ be a path which does not permute the values of f. Study the monodromy of the function h along such paths (considering the homotopy classes of the paths, non permuting the values of f, as forming a subgroup of the fundamental group of the complement to the ramification points of h);
B. Prove that the monodromy group of the function h is solvable, if the monodromy groups of f and of g are solvable;
C. Prove that the monodromy group of every algebraic subfunction of an algebraic function is solvable. [A subfunction is a function whose values (at each argument point) form a subset of the set of the values of the larger function.];
D. Prove that the algebraic function $y = f(x)$, defined over $\mathbb{C}$ by the equation $y^5+xy+1=0$, is not representable as a combination of radicals (and of continuous univalued functions).

Chapter 15: Singularity Theory

1. Prove that any even smooth function of one variable, $f(x) \equiv f(-x)$, $x \in \mathbb{R}$, can be represented as a smooth function of the argument x^2: $f(x) \equiv F(x^2)$.

2 (*Semicubic Cusp*). Let $h : \mathbb{R} \to \mathbb{R}^2$ be a smooth map, whose derivative vanishes at 0. Suppose that h is generic among such applications.
A. Find a coordinate system (x, y) of $\mathbb{R}^2$ in which $h(t) = \big(x(t), y(t)\big)$, where

$$x(t) = t^2 + o(t^2)\ , \qquad y(t) = t^3 + o(t^3) \quad (t \to 0). \tag{1}$$

B. Let h a smooth map satisfying (1). Find a coordinate system (X, Y) of $\mathbb{R}^2$ (in a neighbourhood of the origin) where $h(t) = \big(X(t) = t^2, Y(t) = t^3\big)$.

This means that *all generic cusps are diffeomorphic to the semicubic one.*

3 (*Involute and Fronts*). Let γ be a smooth curve in Euclidean plane $\mathbb{R}^2$, parametrised by the (oriented) arc-length s. Let t be the oriented length of the

segment PQ of the tangent line to γ at the point P, corresponding to the value s of the curve parameter.
A. Calculate the critical points and the critical values of the map $\mathbb{R}^2 \to \mathbb{R}^2$, sending point (s,t) to the point $Q = Q(s,t) \in \mathbb{R}^2$;
B. The points $Q(s,t)$ corresponding to pairs (s,t) such that $s+t$ equals a constant T, form a curve $F_T \subset \mathbb{R}^2$, called "front at time T" or "involute" of the curve γ. Prove that the fronts are orthogonal to their *rays* (the tangent lines PQ of the curve γ) at the points Q;
C. Prove that the front is locally diffeomorphic to a semicubic cusp near any generic point of the curve γ;
D. Investigate the fronts F_T of the cubic parabola ($\gamma : y = x^3$ in Cartesian coordinates, $t = 0$, $s = 0$ at the origin). Draw the fronts F_{-1}, F_0, F_1 and study their singularities. Discuss the stability of the obtained results (with respect to deformations of the curve γ).

4 (*Normal map*). Let M^2 be the space of the normal fibre bundle of the parabola $y = x^2$ in Euclidean plane $\mathbb{R}^2$ (with coordinates (x,y)).
A. Denote $N : M^2 \to \mathbb{R}^2$ the normal map: $N(p) = p + q$ for p a normal vector to the parabola at its point q. That is, the normal map N associates to the vector p (based at q) its endpoint $q + p$.

(i) Find the critical points and the critical values of the map N.

(ii) Draw the variety of the critical values of N and draw the field of the kernels of the derivative map of N (in the chart of M with local coordinates $x(q), |p|$).
B. Let α, β, γ be three germs of smooth curves in M^2, at a critical point of N. Describe the singularities of the curves $N\alpha$, $N\beta$ and $N\gamma$ in the following cases.

(i) The curve α starts at a generic critical point of N.

(ii) The curve β is tangent to the kernel of the derivative of N at a generic critical point of N.

(iii) The curve γ is tangent to the kernel of the derivative of N at a critical point in which the map N is not a folding.

(iv) Find the images of the curves $|p| = \text{const}$ of M under the map N, draw them for all the constant values and describe their singularities (up to diffeomorphisms of the plane).
C. Denote by $F_Q(X) = \|Q - X\|^2$ the family of functions, defined on the parabola, and depending on the parameter $Q \in \mathbb{R}^2$.

(i) Draw the variety of those values of Q for which the function F_Q is not a Morse function; describe the singularities of this variety (up to diffeomorphisms of the plane) and describe the types of critical points of F_Q for these values of Q.

(ii) Draw the variety of the points $Q_* \in \mathbb{R}^2$ for which the function $u(Q) = \min_X F_Q(X)$ is not C^∞ at Q_*. Are the germs of u at these points Q_* mutually R^+-diffeomorphic (up to the addition of constants)?

(*iii*) Is the graph of the (multivalued) function of the point Q, defined as $v(Q) =$ (squared distance from Q to the points of the parabola in which the normal to the parabola contains Q), locally diffeomorphic to the swallowtail ?

2 (*Focal surface*). Find (and draw) the envelope of the family of normal lines to the surface defined by the equation $z = x^2 + 2y^2$ in Euclidean space $\mathbb{R}^3$. Find the number of connected components of the complement to this surface in $\mathbb{R}^3$.

5 (*Caustic and Maxwell stratum*). Let $F(x; a, b) = x^5 - x^3 + ax^2 + bx$ be a family of functions of one real variable x, depending on two real parameters (a, b).
A. Draw the caustic of the family (formed by the parameter values for which the function has a degenerate critical point) and draw the projectively dual curve;
B. Draw the Maxwell stratum of the family (formed by those parameter values for which the function has coinciding critical values at different critical points);
C. Draw the domains of the (a, b)-plane corresponding to LR-stable functions $F(\,\cdot\,; a, b)$ and draw the graphs of the functions corresponding to each domain;
D. Find the triples $(x; a, b)$ for which the germ of the family F at the point $(x; a, b)$ is an R^+-versal deformation of the germ of the function $F(\,\cdot\,; a, b)$ at the point x;
E. Same question, as in item D for the R-versal and V-versal deformations.

6. Consider the family (with 3 real parameters a, b, c) of functions of one real argument x, $f(x) = x^5 + ax^3 + bx^2 + cx$. Find in $\mathbb{R}^3 = \{(a, b, c)\}$ the domain corresponding to the functions with 4 real critical points. Decompose this domain into subsets of functions topologically equivalent to each other. Draw the intersections of these equivalence classes with the plane $a = -1$.

7. Let $f_a = x^3 + axy^2$ be a real function of $(x, y) \in \mathbb{R}^2$.
A. Draw the level lines and the fastest descent directions of f_a for $a = 1,\ 0,\ -1$;
B. Which pairs f_a, f_b are R-equivalent?
C. Find the set A of all the values of the parameter a such that the gradient ideal of f_a has finite codimension. From item D to item H we assume that $a \in A$.
D. Choose a minimal set of additive generators for the vector-space of the local algebra of f_a (which is the quotient-space modulo the gradient ideal) and calculate the multiplication table of this local algebra (for the chosen generators);
E. Calculate the multiplicity of the critical point of f_a;
F. Find all the values of r such that the r-jet of f_a at 0 is sufficient;
G. Find a R^+-versal deformation of f_a;
H. Find the number of connected parts into which the base of the R^+-versal deformation of f_a is decomposed by the union of the caustic and the Maxwell stratum;
I. Take $a = \pm 3$ and let $F_\pm(x, y; \lambda) = f_{\pm 3}(x, y) + \lambda_1(x^2 \mp y^2) + \lambda_2 x + \lambda_3 y$.

(*i*) Find the number of real critical points of the function $F_\pm$ of variables x, y for different values of the parameter λ;

(*ii*) Find the number of non R^+-equivalent germs among the functions of the family F_+ (of the family F_-);

(*iii*) Draw the sections by the planes $\lambda_1 = 1$, 0 and -1 of the caustic and of the Maxwell stratum of the family F_+ (of the family F_-);

(*iv*) Draw the surface formed by those values of the parameter λ for which the function $F_\pm$ has degenerate critical points (*the caustic of the family of functions*).

Chapter 16: Symplectic and Contact Geometry

1 (*Gauss map, Lagrange Singularities and Flattening of Space Curves*). Let γ be an embedded smooth closed curve in Euclidean space $\mathbb{R}^n$.

A. Prove that the normal vectors to the curve γ form an exact Lagrangian submanifold, embedded into the cotangent bundle space $T^*\mathbb{R}^n$.

B. Prove that the oriented normal straight lines of γ form a Lagrangian subvariety of the cotangent fibration space of the sphere of dimension $n-1$. Is this variety (generically) smooth? Is it (generically) exact?

C. The Gauss map Γ of the curve γ sends the spherised normal bundle (of unit normal vectors of γ) to the sphere $\mathbb{S}^{n-1}$, sending each normal vector to its direction. Prove that Γ is a Lagrangian map and find its degree.

D. Find the critical points of the Gauss map Γ for the generic smooth curves in 3-dimensional Euclidean space.

E. Describe the singularities of the critical value sets of the Gauss map Γ for the generic smooth curves in the Euclidean spaces of dimension 3 and 4.

F. Find the maximal number of singularities of the caustics of the Gauss maps Γ for those generic curves in Euclidean 3-space which are close to the standard plane circle (in metric C^3).

2 (*Gauss map, Lagrange Singularities and Parabolic Curve of Surfaces*). Let $z = y^2 + x^2y + x^5$ be the equation of a surface in Euclidean space $\mathbb{R}^3$.

A. Find the critical points of the Gauss map of this surface, compute the kernel of its derivative at each point and draw the caustic curve (formed by the critical values) of the map:

(*i*) at the neighbourhood of the image of $(x = y = z = 0)$;

(*ii*) at other places;

B. Describe the Lagrangian types of the singularities of item A in the Lagrangian singularities hierarchy (A, D, E,...);

C. Describe the singularities of the dual surface of the initial surface, finding their places in the hierarchy of Legendrian singularities of fronts. Then draw the germ of this dual surface at the point that corresponds to $(x = y = z = 0)$ in the original surface and draw the whole dual surface (front).

D. Are the two Lagrangian A_3-singularities of the map of (A) Lagrange-equivalent?

E. Deform the equation into the 1-parameter family $z = y^2+x^2y+x^5+Ax^3$, $A \in \mathbb{R}$.

(*i*) investigate the bifurcation of the caustics of the family, happening near the value $A = 1/40$. Describe the surface formed by the bifurcating caustics in the product of the sphere $\mathbb{S}^2$ with the axis of the parameter A.

(ii) Describe the plane $z = px + qy + r$ by the point $(p, q, r) \in \mathbb{R}^3$. Describe the singularities of the union of the dual surfaces (to our surfaces, depending on A), considering this union as a 3-variety in the 4-space with coordinates $(A; p, q, r)$ (being the moving front in the space-time).

3. Let $S_{u,v}(x) = \cos 2x + u \sin x + v \cos x$ be a 2-parameter family of functions.
(a) Find the Lagrangian surface determined by the generating family $S_{u,v}$ in the space of the cotangent bundle of the plane $\{(u, v)\}$;
(b) Draw the caustic of the Lagrangian surface in the plane $\{(u, v)\}$.
(c) Is this caustic diffeomorphic, near its singular points, to the semicubic cusp?

4. Study the singularities of the dual curve of the cubic parabola $y = x^3$ and draw the curve dual to its perturbation $y = x^3 - ax$.

5. Find the Legendre transforms of the functions $y = x^a$, $a = 2, 3, 4, 5$, and draw the graphs of the transformed functions.

6 (*Legendrian Knots*). Consider the standard contact space $\mathbb{R}^3 = \{x, p, y\}$ equipped with the contact structure $dy = p\, dx$.
A. Calculate the contact self-linking number of the Legendrian curve of the contact space $\mathbb{R}^3$, whose front under the Legendrian fibration $(x, p, y) \mapsto (x, p)$ is the plane curve $(x = \cos t,\ p = \sin 3t)$ (the Legendrian curve is obtained by $y = \int p dx$ up to a shift along y). It is the linking number of this curve with its deformation, deforming the curve at every point in a direction transverse to the contact plane;
B. Find the linking numbers of closed Legendrian curves in the standard contact space $\mathbb{R}^3$ with all the C^1-close closed Legendrian curves.

7 (*Fronts and Optical Lagrangian Singularities*). Let $T^*\mathbb{R}^2 = \{q_1, q_2, p_1, p_2\}$, $\omega = dp_1 \wedge dq_1 + dp_2 \wedge dq_2$ be the standard symplectic space.
A. Find the characteristic curves of hypersurface $p^2 = 1$;
B. Find those characteristics of this hypersurface, which contain the points of the hypersurface where $q_2 = q_1^2$ and $p_1 + 2q_1p_2 = 0$. Is the surface $\mathbb{S}^2$, formed by these characteristics Lagrangian? Is it embedded into $T^*\mathbb{R}^2$?
C. Whatever be the (smooth) submanifold Γ of $\mathbb{R}^n$, prove that the union of the characteristics of the hypersurface $p^2 = 1$ which contain the points $\big(q \in \Gamma, p$ orthogonal to $T_q\Gamma\big)$ is an immersed Lagrangian submanifold of $T^*\mathbb{R}^n$;
D. Find the critical points of the projection $\mathbb{S}^2 \to \mathbb{R}^2$ (of the surface of B) induced by the standard fibration $T^*\mathbb{R}^2 \to \mathbb{R}^2$ of the cotangent bundle, and find the kernels of the derivative of this projection at the critical points;
E. Find the curve formed by the critical values of this projection and the singularities of this caustic;
F. Find and draw the envelope of the projections of the characteristics of item B to the plane (q_1, q_2);

G. Find the points in which the kernels of item D are tangent to the curve of critical points, and study the Lagrangian stability of the singularities of the projection;
H. Given a Lagrangian surface $\mathbb{S}^2$ of $T^*\mathbb{R}^2$, define the function $z : \mathbb{S}^2 \to \mathbb{R}$ by the formula $z = \int p\, dq$. Draw the image F of the map $\mathbb{S}^2 \to \mathbb{R}^3$ whose components are (q_1, q_2, z) for the surface $\mathbb{S}^2$ of item B;
I. At those points in which the surface F is locally the graph of a function $z = u(q_1, q_2)$, calculate $(\partial u/\partial q_1)^2 + (\partial u/\partial q_2)^2$;
J. Draw the level curves $\Gamma_t :\ u(q_1, q_2) = t$ in the plane (q_1, q_2);
K. Find a Legendrian map whose front is the surface F and find its A_2 and A_3 singularities. Investigate the Legendrian stability of these singularities;
L. Find some Legendrian maps L_t whose fronts are the curves Γ_t of item J;
M. Find those values of parameter t for which all the Legendrian singularities of the maps L_t are stable.

8 (*Generating Families and Caustics*). Find the caustics defined by the following generating families and draw these hypersurfaces in the q-spaces:
A. $F(q, x) = x^4 + q_1 x^2 + q_2 x + q_3$;
B. $F(q, x, y) = x^3 + 3xy^2 + q_1 x^2 + q_2 x + q_3 y$;
C. $F(q, x, y) = x^3 - 3xy^2 + q_1 x^2 + q_2 x + q_3 y$;
D. $F(q, x, y) =$ the distance from the point with Cartesian coordinates (q_1, q_2) in Euclidean plane $\mathbb{R}^2$ to the point (x, y) of the ellipse $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = 1$ of this plane.
E. Same family as in D for a (generic) ellipsoid in Euclidean space $\mathbb{R}^3$.

9 (*Contact Elements and* 1*-Jets Manifolds*). Let $J^1(M, \mathbb{R})$ be the manifold of 1-jets of functions $M \to \mathbb{R}$, equipped with its natural contact structure $\alpha = 0$, where $\alpha = dy - p\, dx$. Given a manifold V, denote by $ST^*(V)$ the manifold of *contact elements of* V (affine co-oriented hyperplanes of the tangent spaces of V) equipped with his natural contact structure (the velocity vector of a curve in this space belongs to the contact structure hyperplane if the contact point moves in a direction belonging to the contact element).
A. Are $J^1(\mathbb{S}^1, \mathbb{R})$ and $ST^*(\mathbb{R}^2)$ diffeomorphic? Contactomorphic?
B. Same questions for $J^1(\mathbb{S}^n, \mathbb{R})$ and $ST^*(\mathbb{R}^{n+1})$.

10. Compare the natural symplectic structures of the manifold $T^*\mathbb{S}^{n-1}$ and of the space of the orbits of the Hamilton vector-field of the function $H(p, q) = p^2$ on the hypersurface $p^2 = 1$ of the phase space $\mathbb{R}^{2n} = T^*\mathbb{R}^n$.
Are these two manifolds diffeomorphic? Symplectomorphic?

11. Classify (up to local symplectomorphisms) A. the generic germs of smooth surfaces X^2 in symplectic manifolds M^{2n}; B. the germs of generic smooth surfaces $X^2 \subset M^{2n}$.

12. Same problems for the contactomorphic classification of 2-surfaces in contact 3-manifolds.

13. Same problems for germs of 4-subvarieties of symplectic manifolds M^{2n}.

14. Classify the vector subspaces of the tangent space of a contact manifold M^{2n+1} at point x, up to the x-preserving contactomorphisms of M^{2n+1}.

15. Find the maximal number of vector subspaces $\mathbb{R}^4$ of the symplectic standard vector-space $\mathbb{R}^6$, which are not symplectomorphic to one another.

Chapter 17

1 (*Confocal Quadrics*). Let x, y, z be Cartesian coordinates in Euclidean space $\mathbb{R}^3$ (taken as parameters). For $a < b < c$ consider the equation of one variable $f(\lambda) = 1$, where

$$f = \frac{x^2}{a-\lambda} + \frac{y^2}{b-\lambda} + \frac{z^2}{c-\lambda}\,.$$

A. Prove that, for any point (x, y, z), the roots $\alpha < \beta < \gamma$ of the equation are real and belong to the domains $\alpha \leqslant a \leqslant \beta \leqslant b \leqslant \gamma \leqslant c$;
B. Prove that the 3 surfaces $\alpha(x, y, z) = A$, $\beta(x, y, z) = B$ and $\gamma(x, y, z) = C$ in $\mathbb{R}^3$ are orthogonal at their intersection points;
C. Compute the Betti numbers of the 3 domains of $\mathbb{R}^3$, formed by all the points (x, y, z) for which (i) $\alpha < a$, (ii) $a < \beta < b$, (iii) $b < \gamma < c$;
D. Find a non-constant function u such that the divergence of the gradient of $U(x, y, z) = u(\alpha(x, y, z))$ is everywhere zero in the domain $\{(x, y, z) : \alpha < a\}$;
E. Find a mass distribution producing the gravitational field $\operatorname{grad} U$.

2. Prove that a homogeneous mass distribution on the sphere's surface does not attract interior points of the ball bounded by the sphere.

3. Find an extension of the result of problem 2 to the ellipsoid's surface: which mass distributions along it do not attract the interior points?

4. Extension of the previous fields to the exterior points: find the exterior gravitational field of the mass distributions defined in problems 2 and 3.

5. Generalise problem 1 D to the two domains

$$\begin{array}{ll} \{(x, y, z) : a < \beta(x, y, z) < b\} \;, & \text{div grad } v(\beta(x, y, z)) = 0. \\ \{(x, y, z) : b < \gamma(x, y, z) < c\} \;, & \text{div grad } w(\beta(x, y, z)) = 0. \end{array}$$

[Hint: replace the gravitational fields of the distributions of masses by the magnetic fields of currents of charged particles.]

6. Extend problems 1-5 to Euclidean spaces $\mathbb{R}^n$, $n \neq 3$.

19.2 Miscellaneous Supplementary Problems

1. Find the maximal number of tangencies of a straight line with a smooth generic surface in the space $\mathbb{R}^3$.

2. Compute the integral of the Gauss curvature of the surface

$$x^2(3-x^2)^2 + y^2(3-y^2)^2 + z^2 = 6 \ .$$

3. Compute the dimension D of the space of all the representations of the symmetric group $S(3)$ by unitary operators into $\mathbb{C}^{18}$, equivalent to a given representation. Find the representation for which D is maximal.

2. Calculate the homology groups of the Riemann surfaces $V(A,B) : Y^2 = F(X;A,B)$, depending on the complex parameters A, B, and study the singularities of the differential form dX/Y, restricted to the surfaces V.

3. Study the singularities of the integrals $I(A,B)$ of the form of problem 2 along the cycles of the surfaces $V(A,B)$, considering these integrals as (multivalued) functions of the plane $\mathbb{C}^2$ with coordinates A and B.

Characteristics

A *characteristic* object (of some geometric structure) is an object intrinsically connected to this structure or defined intrinsically in terms of it (being invariant with respect to the structure preserving groups, transformations, diffeomorphisms, etc.).

1. Let $H : \mathbb{R}^{2n} \to \mathbb{R}$ be a smooth function in the standard symplectic space with Darboux coordinates $(p_1,\dots,p_n,q_1,\dots,q_n)$ without critical points : $dH \neq 0$. Find :

A. The characteristic directions of the hypersurface $H = 0$ at each of its points;
B. The characteristic vector-fields of the function H;
C. The characteristic direction of the differential form $p\ dq - H(p,q,t)\ dt$ in the 3-space (with coordinates p,q,t).

2. On the plane with coordinates (x,y), find :

A. The characteristic functions of the vector-field $x\ \partial/\partial x + y\ \partial/\partial y$;
B. The characteristic functions of the vector-field $y\ \partial/\partial x - x\ \partial/\partial y$;
C. The characteristic curves of the vector-field $2x\ \partial/\partial x + 3y\ \partial/\partial y$.

Differential equations – structural stability

1. Consider the system of differential equations $\dfrac{dx}{dt} = y\,,\ \dfrac{dy}{dt} = 2x - x^3\,.$

A. Find the stationary points on the phase plane $\{(x,y)\}$;
B. Find the potential energy (of the given Newton equation);
C. Draw the constant total energy curves on the phase plane;
D. Compute the number of orbits of the phase flow in the constant total energy curve that contains the point $(x = 1, y = 0)$;
E. Find the limits $\lim_{t\to\infty} x(t)$ for the solution with initial conditions $x(0) = 0$, $y(0) = Y$, where (i) $Y = 0$, (ii) $Y = 1/2$, (iii) $Y = 1$, (iv) $Y = -1$, (v) $Y = 2$;
F. Find a local diffeomorphism of the phase plane near its point $(x = 2, y = 0)$

(i) sending the (local) orbits of the phase flow to parallel straight lines,

(ii) sending the phase velocity vector-field to the velocity vector-field of the system $dx/dt = 1$, $dy/dt = 0$;

G. Find the structurally stable stationary points (structural stability is the persistence of the local topological structure of the phase portrait under all sufficiently small deformations of the vector-field, up to the homeomorphisms, close to the identity map);

H. Transform one of the stationary points into an attractor of the deformed vector-field by an arbitrary small deformation of the given field;

I. Find a non-constant solution of the system, tending to finite limits both for $t \to -\infty$ and for $t \to +\infty$. Would the existence of such solution persist under the sufficiently small perturbations of the vector-field?

Hamiltonian Systems

1. Let $H(p,q) = \sin\big(q^2 + (p+q^2)^2\big)$.

A. Which of the stationary points of the Hamilton system, defined by the Hamilton function H, are stable?

B. Find the initial conditions of the periodic solutions of the system; Is the period of these periodic solutions a continuous function of the initial conditions?

D. Construct the action angle coordinates for the Hamilton system;

E. Find the period of the solution of the system with initial conditions $p(0) = \sqrt{\pi}/2$, $q(0) = 0$;

F. Calculate the value of the Poisson bracket of the function H with the function $F(p,q) = (p+q+q^2)^{10}$ at the point $p = 0$, $q = 1$.

2. Calculate the indices of the Hamilton vector-field, with Hamilton function H, at every vanishing points of this field on plane $\mathbb{R}^2$.

Fibred Spaces

1. Calculate the fundamental group and the one-dimensional homology group of the manifold of the length 1 tangent vectors of a genus g surface.

2. For each $k \in \mathbb{Z}$, glue the boundaries of the two direct products $B^2_+ \times \mathbb{S}^1$ and $B^2_- \times \mathbb{S}^1$, where $B^2_+ = \{z \in \mathbb{C} : |z| \leqslant 1\}$, $B^2_- = \{w \in \mathbb{C} : |w| \leqslant 1\}$, $\mathbb{S}^1 = \{\xi = e^{i\varphi}\}$, by the the following identification

$$A_k : \partial(B^2_+ \times \mathbb{S}^1) \to \partial(B^2_- \times \mathbb{S}^1)\ , \quad A_k(z,\xi) = (w = z, \eta = z^k \xi)\ .$$

The resulting 3-manifold M_k is naturally fibred into circles $\mathbb{S}^1$ over the two-sphere: $\mathbb{S}^2 =$ (northern hemisphere B^2_+) $\cup$ (southern hemisphere B^2_-). Given two integers $k \neq \ell$, are the 3-manifolds M_k and M_ℓ homeomorphic?

3 Is the fibration $\det : \mathrm{U}(2) \to \mathbb{S}^1$ homeomorphic to the direct product fibration $\mathbb{S}^1 \times \mathbb{S}^3 \to \mathbb{S}^1$?

4 Consider the sum of the tangent fibration and of the normal fibration of a smooth submanifold V^m of the Euclidean space. Is this sum the trivial fibration of the product $V^m \times \mathbb{R}^n \to V^m$?

5 Consider the 4-space M^4, fibred over the torus T^2 with torical fibres T^2. Suppose that M^4 is a symplectic manifold. Can its Betti number b_1 be odd?

6 Prove that the manifold Λ^3 of the Lagrangian vector subspaces of the standard symplectic space $\mathbb{R}^4$ is fibred over the circle $\mathbb{S}^1$ (with fibres $\mathbb{S}^2$) by the "squared determinant" map $\det^2 : \big(\mathrm{U}(2)/\mathrm{O}(2)\big) \longrightarrow \mathbb{S}^1$. Is this manifold Λ^3 diffeomorphic to the direct product $\mathbb{S}^1 \times \mathbb{S}^2$?

7 Find the mistakes in the two proofs of the triviality of the fibration ($\Lambda^3 \approx \mathbb{S}^1 \times \mathbb{S}^2$) in the textbook "Geometric Asymptotics" by V. Guillemin and S. Sternberg.

Flux of vector fields

1. Find the flux of the vector-field $\operatorname{grad} u$ through the curve $x^2 + y^2 = 1$ in Euclidean plane, with Cartesian coordinates (x, y), for the following functions u:
A. The angle APB, where $[A, B]$ is a segment in the Euclidean plane and P is a point with coordinates (x, y) in the complement the segment $[A, B]$;
B. $u(x, y) = \ln|z - A|$ and $u(x, y) = \operatorname{arctg}\dfrac{y-b}{x-a}$, where $z = x + iy$, $A = a + ib$;
C. The function $u(x, y) = \ln\left|z^5 + \dfrac{z}{2} - \dfrac{1}{10}\right|$;
D. Generalise (a) to Euclidean 3-space replacing $[A, B]$ by an (oriented) plane disc.

2. Compute the flux of the vector-field $\overrightarrow{r}/|\overrightarrow{r}|^3$ (where $\overrightarrow{r} = (x, y, z)$) through the surface $\{(x = (2\cos s) + 1\,,\ y = \sin(2s)\ \ z = \cos t)\}$ in Euclidean space $\mathbb{R}^3$.

3. Is every vector-field v, having zero divergence in $\mathbb{R}^3 \setminus \{0\}$, representable by the formula $v = \operatorname{rot}(w)$? (there exists w ?)

4. For two divergence free vector-fields u and v in Euclidean space $\mathbb{R}^3$ denote by $\{u, v\}$ the vector-field $\operatorname{rot}[u \wedge v]$ (where $[u \wedge v]$ means the vector product).
A. Compute $\big\{\{u, v\}, w\big\} + \big\{\{v, w\}, u\big\} + \big\{\{w, u\}, v\big\}$ (for $\operatorname{div} u = \operatorname{div} v = \operatorname{div} w = 0$);
B. Let u and v be the velocity fields of two flows (i.e., one-parameter groups of diffeomorphisms) of an incompressible fluid, $\{g^t\}$ and $\{h^t\}$. Prove that $\{u, v\} = 0$ if and only if the flows commute: $g^t h^s = h^s g^t$ for any s and t.

19.3 Three Examinations

Examination in Geometry 2005

1. Is the manifold A of affine lines in Euclidean plane $\mathbb{R}^2$ orientable? Compute its fundamental group $\pi_1(A)$ and its homology groups $H_i(A)$.

2 Let B the manifold of affine oriented lines in Euclidean plane $\mathbb{R}^2$. Compute $\pi_1(B)$ and $H_i(B)$.

3 Is the manifold C of oriented lines of the projective space $\mathbb{R}P^3$ orientable? Compute its dimension and its fundamental group $\pi_1(C)$.

4 Same questions for the manifold D of the non-oriented lines of the projective space $\mathbb{R}\mathrm{P}^3$.

5 Same questions for the manifold G of the vector subspaces $\mathbb{R}^2 \subset \mathbb{R}^4$. Is G simply connected?

6 Let Q be the 3-manifold of the nondegenerate quadratic forms on the plane $\mathbb{R}^2$. Compute $\pi_i(Q)$ and $H_i(Q)$.

7 Describe the set V of critical values of the map $f : \mathbb{R}^3 \to \mathbb{R}^3$, sending the point (x, y, z) to $(x^4 + x^2y + zx, y, z)$. Compute the homology groups $H_i(\mathbb{R}^3 \setminus V)$ of the complement to V in space $\mathbb{R}^3$.

8 Let $\varkappa : t \mapsto \big(\cos(2t)(2+\cos(3t)), \sin(2t)\,(2+\cos(3t)), \sin(3t)\big)$ be the "trefoil knot" curve in $\mathbb{R}^3$. Is $\varkappa$ the boundary of any compact oriented smooth surface embedded in space $\mathbb{R}^3$? If it exists, it suffices to draw its image.

Examination in Symplectic Geometry 2002

1 Solve the equation $u_x^2 + 4u_y = 0$ with the boundary condition $u(x, 1) = x^2$.

2 Classify the linear symplectomorphisms $\mathbb{R}^2 \to \mathbb{R}^2$ up to symplectomorphic diffeomorphisms of $\mathbb{R}^2$.

3 Are all the Lagrangian sections of the cotangent bundle fibration $T^*M \to M$ symplectomorphic to each other?

4 The period of the oscillations, with energy $H(p, q) = E$, equals $1 + E$. Find the area bounded by the curve $H = E$ in the phase plane.

5 Let $\mathbb{R}^{2n}$ be the standard symplectic space. Find the number of Lagrangian n-planes among the C_{2n}^n subspaces defined by $x_1 = \cdots = x_n = 0$, where the x_k are some n of the $2n$ Darboux coordinates $(p_1, \ldots, p_n, q_1, \ldots, q_n)$.

6 Is it possible to connect any two points of any connected contact manifold by an integral curve of the contact structure?

7 Is it possible to transport any point of any contact variety along any given smooth curve by a one-parameter family of contactomorphisms? By a flow (that is, by a one-parameter group of contactomorphisms)?

8 One of the eigenvalues of the linearisation of a contactomorphism of $\mathbb{R}^3$ at a fixed point equals $1 + i$. Find the two other eigenvalues.

9 Find all the characteristic directions of a smooth surface of dimension $2n - 1$ in a symplectic manifold of dimension $2n$ (a direction is called *characteristic* for the surface if it is defined by it intrinsically, that is, it remains unchanged under the action of the symplectomorphisms, preserving the surface and the contact point).

10 Find two Lagrangian maps which are diffeomorphic to the map $(x, y) \mapsto (x^3 + xy, y)$, but are not equivalent as Lagrange maps (which are, by definition, the Lagrange submanifold restrictions of the local fibred symplectomorphisms of the cotangent bundle local fibration).

11 Let $z = p + iq$, $w = P + iQ$, $\omega = dp \wedge dq + dP \wedge dQ$, $a \in \mathbb{C}$, $b \in \mathbb{C}$. Are the surfaces $w = az^2$ and $w = bz^2$ symplectomorphic?

12 Can a contact vector-field in $\mathbb{R}^3$ have any attractors?

Examination in Symplectic Geometry 2005

1. Do there exist two non symplectomorphic symplectic structures on $\mathbb{R}^4$?

2. Is it possible to endow the projective space $\mathbb{R}\mathrm{P}^4$ with a symplectic structure?

3. Can two diffeomorphic Lagrangian submanifolds be not symplectomorphic

A. in $\mathbb{R}^2$? B. in the standard symplectic space $\mathbb{R}^4$?

C. in their neighbourhoods (in the symplectic space $\mathbb{R}^4$) ?

4. Does the integral of a symplectic structure ω along a surface, with fixed boundary γ, depend on the surface

A. in the standard symplectic space $\mathbb{R}^4$?

B. in an arbitrary symplectic manifold $(\mathbb{R}^4, \omega)$?

5. Find all the symplectomorphisms of the form $(p, q) \mapsto (P(p, q), q)$ of the standard symplectic space $\mathbb{R}^4$. Do the same for $\mathbb{R}^{2n}$.

6. Does it exist a symplectomorphism $(p, q) \mapsto (P(p, q), Q(q))$ for any diffeomorphism $q \mapsto Q(q)$ of the plane $\mathbb{R}^2$?

8. Find all the Lagrangian submanifolds of the standard symplectic space $\mathbb{R}^4$, belonging to the 3-space $\{q_1 = 0\}$.

9. Find a Lagrangian submanifold L of the standard symplectic space $\mathbb{R}^4$, belonging the hypersurface $\{q_2 = q_1^2\}$ and containing the curve $\{p_1 = p_2 = 0, q_2 = q_1^2\}$.

10. Let $\{g^t\}$ be the flow of the Hamilton system of energy $H = (p_1^2 + p_2^2)/2$ in the standard symplectic space $\mathbb{R}^4$. Find the caustic of the Lagrangian submanifold $g^1 L \subset \mathbb{R}^4$ (for the Lagrangian submanifold L of problem 9): this caustic is a curve of the plane $\mathbb{R}^2$ with coordinates q_1, q_2, whose cotangent bundle is the standard $T^*\mathbb{R}^2 = \mathbb{R}^4$.

To decide later

1. How large can be the number of connected components of a closed smooth curve V^1 in the Euclidean space $\mathbb{R}^N$, whose tangent hyperplanes $\mathbb{R}^{N-1}$ form a variety E^m of given m-volume in the Riemannian manifold of the affine hyperplanes of $\mathbb{R}^N$?

2. Estimate (from above) the Morse number (the minimal number of critical points of a Morse function) of a compact smooth submanifold $V^n \subset \mathbb{R}^N$ in terms of the volume of the variety of its affine tangent hyperplanes.

3. Estimate (from above) the Betti numbers of V^n of problem 2 in terms of its volume and of its curvature.

4*. Let $Y^k(t)$ and Z^ℓ be two compact submanifolds of a compact Riemannian manifold X^m, Y^k depending generically on a parameter $t \in B^p$ (ball of dimension p). Let $A : X^m \to X^m$ be a diffeomorphism, and denote the integer $k + \ell - m \geqslant 0$ by s. Prove that the Betti numbers of subvarieties $U_n^s(t) = (A^n Z^\ell) \cap (Y^k(t))$ are growing with n at most exponentially, $b_*(U_n^s(t)) \leqslant Ce^{\lambda n}$ for almost every value of the parameter t (the exceptional values forming a set of Lebesgue measure 0 in B^p, at least for p which are not too small).

****

1. Let $\gamma(N)$ be a curve of geodesic curvature $Nk(x)$ on a Riemannian surface V whose Gaussian curvature at a point x is $K(x)$. Consider the behaviour of $\gamma(N)$ for $N \to \infty$, for the $k(x) > 0$ case.
Describe the motion along the curve $\gamma(N)$ as a slow evolution of the rotation along a small circle (of radius $1/Nk$), whose centre moves slowly along a level line of the geodesic curvature, $k(x) = c$ (provided that $dk \neq 0$), while in the case $k = \text{const}$ on V the centre moves along the Gaussian curvature level line curve, $K(x) = C$.

****

1. For the differential equation $\dfrac{dx}{dt} = v(x, \varepsilon)$ with $v(x) = v_0(x) + \varepsilon v_1(x)$, denote by $\{g_\varepsilon^t\}$ the flow of the vector-field $v(\,\cdot\,, \varepsilon)$:

$$\frac{\partial g_\varepsilon^t}{\partial t}(x) = v(g_\varepsilon^t(x), \varepsilon) \ , \quad g_\varepsilon^0(x) = x \ .$$

Find the coefficients A_1, A_2 and A_3 of the Taylor series $g_\varepsilon^t(x) = g_0^t(x) + \varepsilon A_1 + \varepsilon^2 A_2 + \varepsilon^3 A_3 + \ldots$.

2. Every diffeomorphism sufficiently close to the identity map belongs to the flow of a suitable vector-field?
Hint: Consider the circle diffeomorphisms $\varphi \mapsto \varphi + \alpha + \beta \sin(n\varphi)$, $\varphi (\mathrm{mod}\ 2\pi) \in \mathbb{S}^1$.

Bibliography

[1] **Alekseev V.B.**, *Abel's Theorem in Problems and Solutions.* Based on the lectures by Professor V.I. Arnold. Kluwer Academic Press, 2004.

[2] **Arnold V.I.**, *On a problem of Liouville, concerning integrable problems of dynamics*, Siberian Math. J. **4**, 471-474 (1963).

[3] **Arnold V.I.**, *Sur une propriété topologique des applications globalement canoniques de la mécanique classique*, C. R. Acad. Sci. Paris **261**, 3719-3722 (1965).

[4] **Arnold V.I.**, *Sur la géométrie différentielle des groups de Lie de dimension infinie et ses applications à l'hydrodynamique des fluids parfaits*, Ann. Inst. Fourier (Grenoble) 1966, **16** N1, 319–361.

[5] **Arnold V.I.**, *On a characteristic class entering in the quantisation conditions.* Func. Anal. and its Appl. **1**:1 (1967) 1-14.

[6] **Arnold V.I.**, *On braids of algebraic functions and cohomologies of the swallowtails.* Russ. Math.Surveys, **23**:4 (1968) 247-248.

[7] **Arnold V.I.**, *Cohomology ring of the died braid group.* Math. Notes, **5**:2 (1969) 227-231.

[8] **Arnold V.I.**, *On some topological invariants of algebraic functions.* Trudy (Proceedings) Moscow Math. Soc. **21** (1970) 27-46.

[9] **Arnold V.I.**, *On matrices depending on parameters.* Russ. Math. Surveys **26**:2 (1971) 101-114

[10] **Arnold V.I.**, *Normal forms for functions near degenerate critical points, the Weyl groups A_k, D_k, E_k and Lagrangian singularities*, Funct. Anal. Appl. **6**:4 (1972) 254-272.

[11] **Arnold V.I.**, *Critical points of smooth functions*, Proceedings of the International Congress of Mathematicians, Vancouver, 1974, Vol. **1**, 19-39.

[12] **Arnold V.I.**, *Ordinary Differential Equations*, Spriger-Verlag, New York Heidelberg Berlin, 1992. (Russian version: Nauka, 1974.)

[13] **Arnold V.I.**, *Wave front evolution and equivariant Morse lemma*, Commun. Pure Appl. Math. 29 No. 6, 1976, 557-582.

[14] **Arnold V.I.**, *Some open problems in the theory of singularities*. Proc. S. L. Sobolev Seminar **1** (1976), Novosibirsk, 5-15. Translated in Singularities, Proc. Symp. Pure Math. **40** AMS, 1983, 57-69.

[15] **Arnold V.I.**, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Spriger-Verlag, New York Heidelberg Berlin, 1983, 1988, 1992. (Russian version: Nauka, 1978, 304 pp.)

[16] **Arnold V.I.**, *On some problems in singularity theory.* in: Geometry and Analysi.s, papera dedicated te the memory of V. K. Patodi, M. F. Atiyah (ed)., Bangalore (1980). Reproduced in the Proceedings of the Indian Acad. Sci. Math Sci. **90** (1981), 1-9.

[17] **Arnold V.I.**, *Magnetic analogues of the theorems of Newton and Ivory.* Uspekhi Mat. Nauk **38**, N. 5 (1983), pp. 145-146.

[18] **Arnold V.I.**, *Some algebro-geometrical aspects of the Newton attraction theory.* Progress in Math., Vol. **36** (1983), pp. 1–4.

[19] **Arnold V.I.**, *Some remarks on the elliptic coordinates.* Notes of the LOMI Seminar, Vol. **133** (1984), pp. 38–50.

[20] **Arnold V.I.**, *On the Newtonian potential of hyperbolic layers.* Selecta Math. Sovietica Vol. **4** (1985), pp. 103-106.

[21] **Arnold V.I.**, *First steps of symplectic topology*, Russ. Math. Surveys **41**:6 (1986) 1-21.

[22] **Arnold V.I.**, *Remarks in quasicrystallic symmetries*, in “Progeress in Chaotic dynamics”, Physica-D, **33**:1 (1988) 21-25.

[23] **Arnold V.I.**, *Spaces of functions with mild singularities.* Funct. Anal. Appl. **23** (3) (1989), 1-10.

[24] **Arnold V.I.**, *Huygens and Barrow, Newton and Hook, Pioneers in mathematical analysis and catastroph theory, from the evolvents to the quasicrystals*, Birkhäuser, 1990, 118 pp.

[25] **Arnold V.I.**, *Bernoulli-Euler updown numbers associated with functions singularities, their combinatorics and arithmetics.* Duke Math. J. **63** (1991), 537-555.

[26] **Arnold V.I.**, *Springer numbers and morsifications spaces.* J. Alg. Geom. **1**:2 (1992) 197-214.

[27] **Arnold V.I.**, *Calculus of snakes and combinatorics of Bernoulli. Euler and Springer numbers of Coxeter groups.* Russ. Math. Surveys **47**:1 (1992) 1-45.

[28] **Arnold V.I.**, *Sur les propriétés topologiques des projections Lagrangiennes en géométrie symplectique des caustiques*, Cahiers de Mathématiques de la decision, CEREMADE, 9320, 14/6/93, 9pp. Rev. Mat. Compl. (1) **8** (1995) 109-119.

[29] **Arnold V.I.**, *On topological properties of Legendre projections in contact geometry of wave fronts*, Algebra i analis (S. Petersburg Math. J.), **6**:3 (1994) 1-16.

[30] **Arnold V.I.**, *Topological invariants of plane curves and caustics*, AMS University Lectures Series, Vol. 5, Providence, R.I., 1994, 62 pp.

[31] **Arnold V.I.**, *On the Number of Flattening Points of Space Curves*, Amer. Math. Soc. Trans. Ser. **171**, 1995, p. 11-22.

[32] **Arnold, V. I.**, *Weak asymptotics of the numbers of solutions of Diophantine equations.* Funct. Anal. Appl. **33**:4 (2000) 292-293.

[33] **Arnold V.I.**, *What is mathematics.* Moscow, MCCME, 2002, 104 pp.

[34] **Arnold V.I.**, *Is mathematics needed at High-schools.* Moscow, MCCME, 2001, 32 pp.

[35] **Arnold V.I.**, *Lobachevsky triangles altitudes theorem and the Jacobi identity in the Lie algebra of quadratic forms on symplectic plane*, J. of Geom. and Phys. **24** (2004) (Russian version: "Matematicheskoe Prosveschenie", Moscow, 2005).

[36] **Arnold V.I.**, *Arnold Problems*, xv+639 pp., Springer, 2004.

[37] **Arnold V.I., Gusein-Zade S.M., Varchenko A.N.** *Singularities of Differentiable Maps V. 1*, Birkhäser, 1985, 382pp.

[38] **Arnold V.I., Givental, A.B.** *Symplectic geometry*, Encyclopædia of Math. Sciences, Dynamical Systems 4 (Springer, New York, 1990), pp. 4-136.

[39] **Arnold V.I., Goryunov V.V., Lyashko O.V., Vassiliev V.A.**, *Singularities 1.* Itogi Nauki and Techn. VINITI, Sovzem. Pro bl. Math. Fund Na.pr. 6 (1988), English translation: Encycl. of Math. Sci. 6, 1992, Springer.

[40] **Arnold V.I., Goryunov V.V., Lyashko O.V., Vassiliev V.A.**, . Singularities-2, Itogi Nauki and Techn. VINITI, Sovzem. Probl. Math. Fund. Napr. 39 (1989), English translation: Encycl. of Math. Sci. 39, 1993, Springer.

[41] **Arnold V.I., Khesin B.A.**, *Topological methods in hydrodynamics*, Springer-Verlag, 1999, 376 pp.

[42] **Atiyah, M**, *New invariants of* 3- *and* 4- *manifolds*, in The mathematical heritage of Hermann Weyl, Durham, N.C., 1987. Sympos. Pure Math., **48** (American Mathematical Society, Providence, RI, 1988), pp. 285-289.

[43] **Audin, M.**, *Cobordisms d'immersions Lagrangiennes et Legendriennes*, Travaux en Cours, 20 (Hermann, 1987).

[44] **Banchoff T., Gaffney T., McCrory**, *Cusps of Gauss mappings*, Research Notes in Math., No. 55, Pitman, Boston (1982).

[45] **Bar-Natan D.**, *Perturbative aspects of the Chern-Simons topological quantumm-field theory.* Ph.D. Thesis, Princeton University, 1991.

[46] **Bar-Natan D.**, *Weights of Feynman Diagrams and the Vassiliev Knot Invariants.* preprint, Princeton University, 1991. It is almost a subset of *On the Vassiliev Knot Invariants*, Topology **34** (1995) 423-472.

[47] **Batyrev, V.**, *Dual polyhedra and mirror symmetry for Calabi-Yau hypersurfaces in toric varieties*, J. Alg. Geom. **3**, (1994).

[48] **Bennequin, D.**, *Entrelacements et equations de Pfaff*, Astérisque 107-108, 83-161 (1983).

[49] **Biran, P.**, *Symplectic packing in dimension* 4, Geom. Funct. Anal. **7** (1997), no. 3, 420–437.

[50] **Birman J.S., Lin X.-S.**, *Knots polyrtomials and Vassiliev invariants*. Invent. Math. **111**:2 (1993) 225-270.

[51] **Blum L., Shub M., Smale S.**, *On a theory of computations over the real numbers: NPcompleteness, recursive functions and universal machines*. Bull. Amer. Math. Soc. **4**:1 (1989) 1-46.

[52] **Blum L., Cucker F., Shub M., Smale S.**, *Complexity and real computations: a manifesto*. Intern. J. of Bifurcation and Chaos **6**:1 (1996) 3-26.

[53] **Borel, A.**, *Sur la cohomologie des espaces fibres principaux et des espaces homogènes de groupes de Lie compacts*. Ann. of Math. **57** (1953) 115-207.

[54] **Brusilovsky, A.D.; Rabinovich, B.I.**, *From* $R-1$ *to* $N-1$, TSNIIMASH, 2005.

[55] **Carlini, F.** *Ricerche sulla convergenza della serie che serve alla soluzione del Problema di Keplero*, Milano 1817; Schumacher Astronomische Nachrichten **28**, 257-270; **30**, 197-254.

[56] **Chekanov, Yu.**, *Legandrova teorija Morsa*, Usp. Mat. Nauk **42**, 139-141 (1987) (Uspekhi are in general translated as Russian Math. Surveys, but some papers contain something new and hence are not translated).

[57] **Chekanov Yu., Pushkar P.**, *Combinatorics of fronts of Legendrian linkings and Arnold's 4-conjectures*, Russian Math. Surveys **60**:1 (2005) 95-149.

[58] **Chern, S. S.; Moser, J. K.**, *Real hypersurfaces in complex manifolds*. Acta Math. **133** (1974), 219-271.

[59] **Chern, S. S.; Moser, J. K.**, *Erratum: "Real hypersurfaces in complex manifolds"*. Acta Math. **150**:3-4 (1983) 297?.

[60] **Conley C.; Zehnder, E.**, *The Birkhoff-Lewis fixed point theorem and a conjecture of V. I. Arnold*, Invent. Math. **73**, 33-49 (1983).

[61] **Eisenbud D.**, *Commutative Algebra with view towards Algebraic Geometry*, Springer-Verlag, 1995, 788pp.

[62] **M. Eisermann**, *Vassiliev Invariants and Poincaré conjecture*. Topology vol. 43, n 5 (2004) 1211-1230.

[63] **Ekeland I.; Hofer, H.**, *Symplectic topology and Hamiltonian dynamics, I,II*, Math. Z. **200**, 355-378 (1989); **203**, 553-568 (1990).

[64] **Eliashberg Y.; Gromov, M.**, *Convex symplectic manifolds*, Proc. Symp. Pure Math **52**, 135-162 (1991);

[65] **Eliashberg, Y.**, *Contact* 3*-manifolds twenty years since J. Martinet's work*, Annales Inst. Fourier **42**, 165-192 (1992).

[66] **Eliashberg, Y.**, *Legendrian and transversal knots in tight contact* 3*-manifolds*, Topological Methods in Modern Mathematics, J. Milnor's 60th birthday volume (Publish or Perish, Houston, 1993), pp. 171-193.

[67] **Eliashberg, Y.; Polterovich, L.**, *Unknottedness of Lagrangian surfaces in symplectic* 4*-manifolds*, Int. Math. Res. Notices, **11**, 295-301 (1993)

[68] **Eliashberg, Y.**, *Classification of contact structures on* $\mathbb{R}^3$, Int. Math. Res. Notices, **3** (1993) 87-91.

[69] **Feynman, R. P.**, *QED: The Strange Theory of Light and Matter*, Princeton University Press, Princeton, 1988.

[70] **Floer, A.**, *Proof of the Arnold conjecture and generalizations to certain Kahler manifolds*, Duke Math. J. **53**, 1-32 (1986).

[71] **Floer, A.**, *An instanton invariant for* 3*-manifolds*, Commun. Math. Phys. **118**, 215-240 (1988).

[72] **Fortune, B.; Weinstein, A.**, *A symplectic fixed point theorem for complex projective spaces*, Bull. Am. Math. Soc. **12**, 128-130 (1985).

[73] **Fulton, W.; Harris, J.**, *Representation theory*, Springer, 1991.

[74] **Fuks D.B.**, *Cohoology of braid groups* mod 2. Func. Anal. and its Appl. **4**:2 (1970) 62–73.

[75] **Gray, J.W.**, *Some global properties of contact structures*, Ann. Math. **2**, 421-450 (1959).

[76] **Gromov, M.**, *Pseudo-holomorphic curves in symplectic manifolds*, Invent. Math. **82**, 307-347 (1985).

[77] **Ginzburg, V.L.**, *Calculation of contact and symplectic cobordism groups*, Topology **31**, 757-762 (1992).

[78] **Givental, A.B.**, *Polynomial electrostatic potentials* (in Russian), Uspekhi Mat. Nauk **39** N.5, 253–254 (1984).

[79] **Givental, A.B.**, *Lagrangian imbeddings of surfaces and the open Whitney umbrella*, Funct. Anal. Appl. **20**, 35-41 (1986).

[80] **Givental, A.B.**, *Singular Lagrange varieties and their Lagrange mappings*, in Itogi Nauki VINITI **33**, 55-112 (1988), Transl. J. Sov. Math. **52**, 3246-3278 (1990).

[81] **Givental, A.B.**, *Nonlinear generalization of the Maslov index*, Singularity Theory and its Applications, edited by V.I. Arnold, Adv. Sov. Math. 1 (American Mathematical Society, Providence, RI, 1990), pp. 71-103.

[82] **Givental, A.B.**, *A symplectic fixed point theorem for toric manifolds*, in Progress in Mathematics (Floer Memorial Volume, Birkhaeuser, 1994).

[83] **Guieu L., Mourre E., Ovsienko V.Yu.**, *Theorem on six vertices of a plane curve via the Sturm theory*, Arnold–Gelfand mathmatical seminars, V. I. Arnold, I.M. Gelfand, M. Smirnov, V.S. Retakh Editors, Birkhäuser Boston, (1997) p.257-266.

[84] **Hofer, H.**, *Symplectic capacities*, in Durham Conference, edited by Donaldson and Thomas (London Mathematical Society, 1992).

[85] **Hurwitz, A.**, *Über die Fourierschen Konstanten integrierbarer Funktionen*, Math. Ann. 57, (1903) p.425-446.

[86] **Kazarian M. E.**, *Multisingularities, cobordisms and enumerative geometry.* (in Russian) Uspekhi Mat. Nauk **58** (2003), no. 4(352), 29-88; English translation: Russian Math. Surveys **58** (2003), no. 4, 665-724.

[87] **Khovansky A.G.**, *Geometry of formulas.* Math. Physics review **4** (1984) 1-92.

[88] **Khovansky A.G.**, *Newton polyhedra.* J. of Soviet Math. **27** (1984), 2812-2833.

[89] **Khovansky A.G.**, *Fewnomials*, PHASIS, Moscow, xii+217 pp., 1997.

[90] **Klein F.**, *Elementary Mathematics From an Advanced Standpoint: Geometry*, Third Ed. Dover publ. Inc., N.Y., 1948.

[91] **Knizhnik V.G., Zamolodchikov A.B.**, *Current algebra and WessÂZumino models in two dimensions.* Nucl. Phys. **B247** (1984), 83-103.

[92] **Kneissler J.**, *The number of primitive Vassiliev invariants up to degree twelve.* ArXiv 1997.

[93] **Kontsevich M.**, *Integrals representing Vassiliev's knot invariants.* Lectures at Bonn MPI, February-March 1991.

[94] **Lauenbach F.; Sikorav, J.-C**, *Persistence d'intersections avec la section nulle au cours d'une isotopie hamiltonienne dans un fibré cotangent*, Invent. Math. **82**, 349-358 (1985).

[95] **Leray J.**, *Sur un théorème d'Arnold* Notes du séminaire (1966).

[96] **Levelt Sengers J.**, *How fluids unmix: Discoveries by the School of Van der Waals and Kamerlingh Onnes*, KNAW, Amsterdam 2002.

[97] **Lin X.-S.**, *Vertex Models. Quantum Groups and Vassiliev knot invariants.* preprint*, Columbia University 19 (1991).

[98] **Martinet, J.**, *Formes de contact sur les variétés de dimension* 3, Lecture Notes in Mathematics, **209**, 142-163 (1971).

[99] **McDuff D., Traynor, L.** (unpublished).

[100] **McDuff D., Polterovich, L.**, *Symplectic packing and algebraic geometry*, Invent. Math. (3) **115** (1994), 405-434.

*FIND!!

[101] **Maslov, V.P.**, *Théorie des perturbations et méthodes asymptotiques*, thesis, Moscow State Univ., 1965 (Dunod, Paris, 1972).

[102] **Milnor J.**, *Morse Theory.* Princeton Univ. Press, New Jersey, 1969.

[103] **Morin B.**, *Formes canoniques des singularités d'une application différentiable.* Compt. Rend. Acad. Sci. Paris, **260** (1965) 5662-5665, 6503-6506;

[104] **Mukhopadhyaya S.**, *New Methods in the Geometry of a Plane Arc I*, Bull. Calcutta Math. Soc. **1**, 1909, p.31–37.

[105] **Newmann W., Wahl, J.**, *Casson invariant of links of singularities*, Comment. Math. Helv. **65**, 58-78 (1990).

[106] **Orlik P., Terao H.**, *Arrangements of hyperplanes.* Springer-Verlag, Grundlehren der Mathematischen Wissenschaften, vol. 300, 1991.

[107] **Papelier G.**, *Exercises de géométrie moderne*, Tome IV (Poles, polaires, plans polaires) Librairie Vuibert, Paris, 1926.

[108] **Rabinowitz, P.**, *Critical points of indefinite functionals and periodic solutions of differential equations*, in Proceedings of the International Congress of Mathematicians, Helsinki 1978 (Acad. Sci. Fennica, Helsinki, 1980), pp. 791-796.

[109] **Saeki O., Yamamoto T.**, *Singular fibers of stable maps and signatures of 4-manifolds*, Geom. Topol. **10** (2006) 359-399.

[110] **Salmon G.**, *A treatise on the analytic geometry of three dimensions.* Vol. 2. Chelsea, N.Y. (1965) 344 p.

[111] **Scherbak, I.G.** *Focal set of a surface with boundary and caustics of groups generated by reflections B_k, C_k, F_4*, Funct. Anal. Appl. **18**, 84-85 (1984).

[112] **Scherbak, O.P.**, *Wave fronts and reflection groups*, Russian Math. Surveys **43**, 149-194 (1988).

[113] **Schwartz A.S.**, *Genus of fibration.* Trudy (Proceedings) Moscow Math. Soc. **10** (1961) 217–272.

[114] **Segal G.**, *Configuration spaces and iterated loop-spaces.* Invent. Math. **21**:3 (1973) 213–221.

[115] **Seade Kuri J. A.**, *On the topology of isolated singularities in analytic spaces*, monograph. Series "Progress in Mathematics" **241** (2005) Birkhäuser, 255 pp.

[116] **Shmelev A.S.**, *On differential invariants of some differential-geometric structures*, Proc. of Steklov Inst. of Math., **209** (1995) 203-234.

[117] **Shapiro, B.Z.; Vainstein, A.D.**, *Higher-dimensional analogs of the theorem of Newton and Ivory.* Funct. Anal. Appl., **19** (1985), 17–20.

[118] **Siegel, C.L.**, *Symplectic geometry*, Am. J. Math. **65**, 1 (1943).

[119] **Smale S.S.**, *On the topology of algorithms.* J. of Complexity **3**:2 (1987) 81-89.

[120] **Sturm J.C.F.**, *Sur une classe d'équations différentielles du second ordre*, J. Math. Pures Appl. 1, (1836) 373-444.

[121] **Sylvester J.J.**, *Note on the spherical harmonics*. Philosophical Magazine II (1876) 291-307. The collected Mathematical Papers of J.J. Sylvester Vol. III,Cambridge (1909) 37-51.

[122] **Tabachnikov S.L.** *Around four vertices*, Russ. Math. Surveys **45**, 229-230 (1990).

[123] **Thomson W.**, *On Vortex Atoms*, Proceedings of the Royal Society of Edinburgh. **6** (1867) 94-105.

[124] **Tresse A.**, *Sur les invariants différentielles des groupes continues de transformations*, Acta Mathematica **18** (1894) 1-88.

[125] **Tresse A.**, *Determination des invariants ponctuels de l'équation différentielle de second ordre*, Leipzig, 1896.

[126] **Uribe-Vargas R.**, *A Projective Invariant for Swallowtails and Godrons, and Global Theorems on the Flecnodal Curve*. Moscow Mathematical Journal, 6:4, (2006) 731-768.

[127] **Vaninsky K.L.**, *Equations of Camassa-Holm type and Jacobi ellipsoidal coordinates*. Comm. Pure Appl. Math. **58**, no. 9, (2005), 1149-1187.

[128] **Vassiliev V.A.**, *Lagrange and Legendre characteristic classes* (Gordon and Breach, New York, 1988).

[129] **Vassiliev V.A.**, *Stable cohomology of the complements to the discriminants of the singularities of smooth functions*. Itogi Nauki and Techn. VINITI, Prob. Math., Noveishie Dostijenia **33** (1988) 3-29, Moscow. English translation: J. of Sov. Math. (1990).

[130] **Vassiliev V.A.**, *Topological complexity of algorithms of approximated computing of roots of systems of polynomial equations*. Algebra and Analysis **1**:6 (1989) 98-113.

[131] **Vassiliev V.A.**, *Topology of the spaces of functions without complicated singularities*. Funct. Anal. Appl. **23** (1989), 24-36.

[132] **Vassiliev V.A.**, *Cohomology of knot spaces*, in: Theory of Singularities and its Applications, V.I. Arnold (ed.), Advances in Soviet Mathematics, AMS 1 (1990), 23-70.

[133] **Vassiliev V.A.**, *Topology of complements to discriminants and loop spaces*. in: Theory of Sirgularities and its Applications, V.I. Arnold (ed)., Advances in Soviet Mathematics 1, AMS, 1990, 9-21.

[134] **Vassiliev V.A.**, *Complements of discriminants of smooth maps: topology and applications*. Transl. of Math. Monographs **98** AMS, Providence, RI (1992) vi+208 pp. ISBN: 0-8218-4555-1. (New Russian vesion: Moscow, Phasis, 1997.)

[135] **Viterbo, C.**, *Symplectic topology as the geometry of generating functions*, Math. Ann. **292**, 685-710 (1992).

[136] **Wassermann, G.**, *Stability of unfoldings in space and time*, Acta Math. **135**, 57-128 (1975) and (r, s)*-stable unfoldings and catastroph theory*; in Lecture Notes in Math. **525**, 253-262 Springer-Verlag (1976).

[137] **Weinstein, A.**, *Lectures on Symplectic Manifolds*, Reg. Conf. Ser. Math. 29 (American Mathematical Society, Providence, 1997).

[138] **Weyl, H.**, *The Classical Groups, their Invariants and Representations.* Princeton University Press, Princeton, NJ, 1939.

Index