Analytic and Differential geometry

Mikhail G. Katz*

Department of Mathematics, Bar Ilan University, Ramat Gan 52900 Israel
Email address: katzmik@math.biu.ac.il

*Supported by the Israel Science Foundation (grants no. 620/00-10.0 and 84/03).

Abstract. We start with analytic geometry and the theory of conic sections. Then we treat the classical topics in differential geometry such as the geodesic equation and Gaussian curvature. Then we prove Gauss's *theorema egregium* and introduce the abstract viewpoint of modern differential geometry. See `http://u.math.biu.ac.il/~katzmik/88-201.html`

Contents

CHAPTER 1

Analytic geometry

(1) At `http://u.math.biu.ac.il/~katzmik/88-201.html` you will find the course site.
(2) There you will find choveret in English as well as tirgul notes by Atia in Hebrew.
(3) The final exam is 90% of the grade, bochan and targil 10%.
(4) The homework assignments can be found at the course site.
(5) Feel free to ask questions via email: katzmik@math.biu.ac.il

After dealing with classical geometric preliminaries including the *theorema egregium* of Gauss, we present new geometric inequalities on Riemann surfaces, as well as their higher dimensional generalisations.

We will first review some familiar objects from classical geometry and try to point out the connection with important themes in modern mathematics.

1.1. Circle, sphere, great circle distance

Definition 1.1.1. The unit circle S^1 in the plane is the locus of the equation

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

in the (x, y)-plane.

The circle solves the isoperimetic problem in the plane. Namely, consider simple (non-self-intersecting) closed curves of equal perimeter, for instance a polygon.[1] [2] [3]

---

[1]metzula

[2]Among all such curves, the circle is the curve that encloses the largest area. In other words, the circle satisfies the boundary case of equality in the following inequality, known as the *isoperimetric inequality*. Let L be the length of the Jordan curve and A the area of the finite region bounded by the curve.

Theorem. Every Jordan curve in the plane satisfies the inequality $\left(\frac{L}{2\pi}\right)^2 - \frac{A}{\pi} \geq 0$, with equality if and only if the curve is a round circle.

[3]The round circle is the subject of Gromov's filling area conjecture. The *Riemannian circle* of length 2π is a great circle of the unit sphere, equipped with the great-circle distance. The emphasis is on the fact that the distance is measured along arcs rather than chords (straight line intervals). For all the apparent

Definition 1.1.2. The 2-sphere S^2 is a surface that is the collection of unit vectors in 3-sphace:

$$S^2 = \left\{ (x, y, z) \in \mathbb{R}^3 \colon x^2 + y^2 + z^2 = 1 \right\}.$$

Definition 1.1.3. A *great circle* of S^2 is the intersection of the sphere with a plane passing through the origin.

Example 1.1.4. The *equator* is an example of a great circle.

Definition 1.1.5. Let $p, q \in S^2$. The *great circle distance* $d(p, q)$ on S^2 is the distance measured along the arcs of great circles connecting a pair of points $p, q \in S^2$:

$$d(p, q) = \arccos(p \cdot q),$$

where $p \cdot q$ is the usual dot product in Euclidean space.

Remark 1.1.6. The distance between a pair of points $p, q \in S^2$ is the length of the smaller of the two arcs of the great circle passing through p and q. In Theorem 6.10.10 we will prove that this arc is the minimal distance between p and q among all curves on S^2 joining p and q. In other words, the geodesic arc is length-minimizing among all paths between a pair of points on the sphere.

Remark 1.1.7. Key concepts of differential geometry that we hope to clarify in our course are the notions of

(1) *geodesic curve* and
(2) *curvature.*

Remark 1.1.8. In relativity theory, one uses a framework similar to classical differential geometry, with a technical difference having to do with the basic quadratic form being used. Nonetheless, some of the key concepts, such as geodesic and curvature, are common to both approaches. In the first approximation, one can think of relativity theory as the study of 4-manifolds with a choice of a

"light cone"[4]

---

simplicity of the the Riemannian circle, it turns out that it is the subject of a still-unsolved conjecture of Gromov's, namely the filling area conjecture. A surface with a single boundary circle will be called a *filling* of that circle. We now consider fillings of the Riemannian circles such that the ambient distance does not diminish the great-circle distance (in particular, filling by the flat unit disk is not allowed). M. Gromov conjectured that *Among all fillings of the Riemannian circles by a surface, the hemisphere is the one of least area.* Partial progress was obtained in [**BCIK05**].

[4]konus ha'or

at every point. Einstein gave a strong impetus to the development of differential geometry, as a tool in studying general relativity. We will systematically use Einstein's summation convention; see Section 1.3. See Sachs and Wu [**Sachs**]. More sources can be found at `https://mathoverflow.net/questions/217565`.

1.2. Linear algebra, index notation

Linear algebra provides an indispensable foundation for our subject. Key concepts here are linear map and bilinear form (see Remark 1.2.3). It is important to develop a facility with Einstein summation convention (treated in detail in Section 1.3). This notation will be exploited throughout the course.

Let $\mathbb{R}^n$ denote the Euclidean n-space. Its vectors will be denoted

$$v, w \in \mathbb{R}^n.$$

DEFINITION 1.2.1. A vector in $\mathbb{R}^n$ is a *column* vector v. To obtain a row vector we take the transpose, v^t.

DEFINITION 1.2.2. Let $M_{n,n}(\mathbb{R})$ be the space of n by n matrices.

Let $B \in M_{n,n}(\mathbb{R})$ be an n by n matrix.

REMARK 1.2.3 (Two ways). There are two ways of viewing a matrix B, either as

(1) a linear map (see Definition 1.2.4) or
(2) a bilinear form (see Section 1.5).

Developing suitable notation to capture this distinction helps simplify differential-geometric formulas down to readable size, and also to motivate the important distinction between a vector and a covector.

DEFINITION 1.2.4 (Matrix as a bilinear form $B(v, w)$). Consider a bilinear form

$$B(v, w) \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \tag{1.2.1}$$

sending the pair of vectors (v, w) to the real number

$$v^t B w. \tag{1.2.2}$$

Here v^t is the row vector given by transpose of v. We write

$$B = (b_{ij})_{i=1,\dots,n;\, j=1,\dots,n}$$

so that b_{ij} is an entry while (b_{ij}), with parentheses, denotes the matrix itself.

EXAMPLE 1.2.5. In the 2 by 2 case, we have

$$v = \begin{pmatrix} v^1 \\ v^2 \end{pmatrix}, w = \begin{pmatrix} w^1 \\ w^2 \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}.$$

Then

$$Bw = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} = \begin{pmatrix} b_{11}w^1 + b_{12}w^2 \\ b_{21}w^1 + b_{22}w^2 \end{pmatrix}.$$

Note that the transpose $v^t = (v^1 \ v^2)$ is a row vector. We therefore calculate the product (1.2.2) to obtain

$$\begin{aligned} B(v, w) &= v^t Bw \\ &= (v^1 \ v^2) \begin{pmatrix} b_{11}w^1 + b_{12}w^2 \\ b_{21}w^1 + b_{22}w^2 \end{pmatrix} \\ &= b_{11}v^1w^1 + b_{12}v^1w^2 + b_{21}v^2w^1 + b_{22}v^2w^2, \end{aligned}$$

and hence $v^t Bw = \sum_{i=1}^{2} \sum_{j=1}^{2} b_{ij}v^iw^j$.

We would like to simplify this notation by suppressing the summation symbols "Σ". The details appear in Section 1.3.

1.3. Einstein summation convention

The following useful notational device was originally introduced by Albert Einstein.

DEFINITION 1.3.1. The rule is that whenever a product contains a symbol with a certain lower index and another symbol with the *same* upper index, take summation over this repeated index, even though the summation symbol Σ is not present.

EXAMPLE 1.3.2. Using this notation, the bilinear form (1.2.1) defined by the matrix B can be written as follows:

$$B(v, w) = b_{ij}v^iw^j,$$

with implied summation over both indices i and j.

To avoid any risk of ambiguity when using the Einstein summation convention, we use underlining as follows. When we wish to consider a specific term such as a_i times v^i rather than the sum over a dummy index i, we use the underline notation as follows.

DEFINITION 1.3.3. The expression

$$a_{\underline{i}}v^{\underline{i}}$$

denotes a specific term with specific index value i (rather than summation over a dummy index i).

1.4. Symmetric matrices, quadratic forms, polarisation

DEFINITION 1.4.1. Let B be a *symmetric* matrix: $B^t = B$. The *associated quadratic form* Q is a quadratic form associated with a bilinear form $B(v, w)$ by the following rule:

$$Q(v) = B(v, v)$$

for all vectors v.

Let $(e_i)_{i=1,\ldots,n} = (e_1, \ldots, e_n)$ be the standard basis of $\mathbb{R}^n$. Given a vector $v \in \mathbb{R}^n$, we write it as $v = v^i e_i$. Each of the components v^i is a real number.

LEMMA 1.4.2. *Given a symmetric bilinear form as above, the associated quadratic form Q satisfies*

$$Q(v) = b_{ij} v^i v^j$$

in terms of the Einstein summation convention.

PROOF. To compute $Q(v)$, we must introduce an extra index j and use it for the second occurrence of the vector v:

$$Q(v) = B(v, v) = B(v^i e_i, v^j e_j) = B(e_i, e_j) v^i v^j = b_{ij} v^i v^j, \qquad (1.4.1)$$

proving the lemma. $\square$

THEOREM 1.4.3. *The* polarisation formula *asserts that*

$$B(v, w) = \frac{1}{4}(Q(v + w) - Q(v - w)).$$

The polarisation formula allows one to reconstruct the symmetric bilinear form from the quadratic form.[5] For an application, see Section 13.10.

1.5. Matrix as a linear map

Given a real $n \times n$ matrix B, consider the associated map

$$B_{\mathbb{R}} \colon \mathbb{R}^n \to \mathbb{R}^n, \quad v \mapsto Bv.$$

In order to distinguish this case from the case of the bilinear form, we will raise the first index of the matrix coefficients. Thus, we write B as

$$B = (b^i{}_j)_{i=1,\ldots,n;\, j=1,\ldots,n}$$

where it is important to s t a g g e r the indices as follows.

[5] This is true whenever the characteristic of the background field is not 2. Our base field $\mathbb{R}$ has characteristic 0 and therefore the formula applies in this case.

Definition 1.5.1. Staggering the indices meaning that we do **not** place j under i as in

$$b^i_j,$$

but, rather, leave a blank space (in the place were j used to be), as in

$$b^i{}_j.$$

Consider vectors $v = (v^j)_{j=1,\ldots,n}$ and $w = (w^i)_{i=1,\ldots,n}$. Then the equation $w = Bv$ can be written as a system of n scalar equations,

$$w^i = b^i{}_j v^j \text{ for } i = 1, \ldots, n$$

using the Einstein summation convention (here the repeated index is j).

Definition 1.5.2. The formula for the trace $Tr(B) = b^1{}_1 + b^2{}_2 + \cdots + b^n{}_n$ in Einstein notation becomes

$$Tr(B) = b^i{}_i$$

(here the repeated index is i).

Remark 1.5.3. If we wish to specify the i-th diagonal coefficient of our matrix B we use the underline notation

$$b^{\underline{i}}_{\underline{i}}$$

as in Definition 1.3.3 (so as to avoid ambiguity).

1.6. Symmetrisation and antisymmetrisation

For the purposes of dealing with symmetrisation and antisymmetrisation of a matrix, it is convenient to return to the framework of bilinear forms (i.e., both indices are lower indices).

Definition 1.6.1. The transpose B^t of a matrix $B = (b_{ij})$ is the matrix whose (i, j)-th coefficient is b_{ji}.

Remark 1.6.2 (Reflection in diagonal). Geometrically the passage from B to B^t corresponds to reflection in the diagonal of the matrix B.

Definition 1.6.3. Let $B = (b_{ij})$. Its *symmetric part* S is by definition

$$S = \frac{B + B^t}{2} = \left(\tfrac{b_{ij}+b_{ji}}{2}\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}},$$

while the antisymmetric (or skew-symmetric) part A is

$$A = \frac{B - B^t}{2} = \left(\frac{b_{ij} - b_{ji}}{2}\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}.$$

Theorem 1.6.4. *We have $B = S + A$.*

Another useful notation is that of symmetrisation and antisymmetrisation at the level of indices.

DEFINITION 1.6.5. Symmetrisation is defined by setting

$$b_{\{ij\}} = \frac{1}{2}(b_{ij} + b_{ji}) \tag{1.6.1}$$

with curly braces,[6] and antisymmetrisation by setting

$$b_{[ij]} = \frac{1}{2}(b_{ij} - b_{ji}). \tag{1.6.2}$$

LEMMA 1.6.6. *A matrix $B = (b_{ij})$ is symmetric if and only if for all indices i and j one has $b_{[ij]} = 0$.*

PROOF. We have $b_{[ij]} = \frac{1}{2}(b_{ij} - b_{ji}) = 0$ since symmetry of B means $b_{ij} = b_{ji}$. □

1.7. Matrix multiplication in index notation

The usual way to define matrix multiplication is as follows. A triple of matrices $A = (a_{ij})$, $B = (b_{ij})$, and $C = (c_{ij})$ satisfy the product relation $C = AB$ if, introducing an additional summation index k (*cf.* formula (1.4.1)), we have the relation $c_{ij} = \sum_k a_{ik} b_{kj}$.

EXAMPLE 1.7.1 (Skew-symmetrisation of matrix product). By commutativity of multiplication of real numbers, we have

$$a_{ik} b_{kj} = b_{kj} a_{ik}.$$

Then the coefficients $c_{[ij]}$ of the skew-symmetrisation of the matrix $C = AB$ satisfy

$$c_{[ij]} = \sum_k b_{k[j} a_{i]k}.$$

Here by definition

$$b_{k[j} a_{i]k} = \frac{1}{2}(b_{kj} a_{ik} - b_{ki} a_{jk}).$$

REMARK 1.7.2. This notational device will be particularly useful in writing down the *theorema egregium* of Gauss (see Section 10.10).

EXAMPLE 1.7.3. Examples of symmetrisation and antisymmetrisation notation:

- See Section 6.3, where we will use formulas of type

$$g_{mj} \Gamma^m_{ik} + g_{mi} \Gamma^m_{jk} = 2 g_{m\{j} \Gamma^m_{i\}k};$$

[6] Sograyim mesulsalim or metultalim

- See Section 9.5 for $L^i{}_{[j}L^k{}_{\ell]}$;
- See Section 10.9 for $\Gamma^k_{i[j}\Gamma^n_{\ell]m}$.

Recall that a real matrix A naturally defines an endomorphism $A_{\mathbb{R}} \colon \mathbb{R}^n \to \mathbb{R}^n$.

THEOREM 1.7.4. *If $A = (a^i{}_j)$, $B = (b^i{}_j)$, and $C = (c^i{}_j)$ then the product relation $C = AB$ corresponding to the composition of linear maps $C_{\mathbb{R}} = A_{\mathbb{R}} \circ B_{\mathbb{R}}$ simplifies to the relation*

$$c^i{}_j = a^i{}_k b^k{}_j \quad \textit{for all } i, j. \tag{1.7.1}$$

The proof is immediate from the definition of matrix multiplication.

REMARK 1.7.5. The index notation we have described reflects the fact that the natural products of matrices are the ones which correspond to composition of maps.

1.8. Two types of indices: summation index and free index

In expressions of type (1.7.1) it is important to distinguish between two types of indices: a free index or an summation index.

DEFINITION 1.8.1. An index appearing both as a subscript and a superscript is called an *summation index*. The remaining indices are called *free*.

A summation index is often referred to as a *dummy index* in the literature.

EXAMPLE 1.8.2. In formula (1.7.1) the index k is a summation index whereas indices i and j are free indices.

1.9. Kronecker delta and the inverse matrix

The Kronecker delta function $\delta^i{}_j$ defined as follows.

DEFINITION 1.9.1. The expression

$$\delta^i{}_j = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases}$$

is called the Kronecker delta function.

Consider a matrix $B = (b_{ij})$.

DEFINITION 1.9.2. The inverse matrix B^{-1} is the matrix

$$B^{-1} = (b^{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$$

where both indices have been raised.

The identity matrix is denoted $I = (\delta^i{}_j)$. Then the equation $B^{-1}B = I$ becomes

$$b^{ik} b_{kj} = \delta^i{}_j \quad \text{for all} \quad i, j \tag{1.9.1}$$

in Einstein notation with repeated index k.

REMARK 1.9.3. In (1.9.1) the index k is a summation index whereas i and j are free indices.

EXAMPLE 1.9.4. The identity endomorphism $I = (\delta^i{}_j)$ by definition satisfies $AI = A = IA$ for all endomorphisms $A = (a^i{}_j)$, or equivalently

$$a^i{}_j \delta^j{}_k = a^i{}_k = \delta^i{}_j a^j{}_k, \tag{1.9.2}$$

using the Einstein summation convention.

REMARK 1.9.5. In expression (1.9.2) the index j is a summation index whereas indices i and k are free indices.

EXAMPLE 1.9.6. Let δ^i_j be the Kronecker delta function on $\mathbb{R}^n$, where $i, j = 1, \ldots, n$. We view the Kronecker delta function as a linear transformation $\mathbb{R}^n \to \mathbb{R}^n$.

(1) Simplify the expression $\delta^i{}_j \delta^j{}_k$ paying attention to which the summation indices are;
(2) Simplify the expression $\delta^i{}_j \delta^j{}_i$ paying attention to which the summation indices are.

REMARK 1.9.7 (Preservation of free indices). When simplifying an expression involving free and summation indices, it is important to check that the final expression has the same free indices as the original expression. It is a common mistake on exams that can easily be avoided by a simple check.

1.10. Vector product

We briefly review the following material from linear algebra.

DEFINITION 1.10.1. Given a pair of vectors $v = v^i e_i$ and $w = w^j e_j$ in $\mathbb{R}^3$, their *vector product* is a vector $v \times w \in \mathbb{R}^3$ satisfying one of the following two equivalent conditions.

(1) (algebraic) Let $i = e_1$, $j = e_2$, $k = e_3$. We have $v \times w = \det \begin{bmatrix} i & j & k \\ v^1 & v^2 & v^3 \\ w^1 & w^2 & w^3 \end{bmatrix}$, in other words,

$$v \times w = (v^2 w^3 - v^3 w^2) i - (v^1 w^3 - v^3 w^1) j + (v^1 w^2 - v^2 w^1) k.$$

(2) (geometric) the vector $v \times w$ is perpendicular to both v and w, of length equal to the area of the parallelogram spanned by the two vectors, and furthermore satisfying the right hand rule,[7] meaning that the 3 by 3 matrix formed by the three vectors v, w, and $v \times w$ has positive determinant.

THEOREM 1.10.2. *We have an identity*

$$a \times (b \times c) = (a \cdot c)b - (a \cdot b)c \tag{1.10.1}$$

for every triple of vectors a, b, c in $\mathbb{R}^3$.

Note that both sides of (1.10.1) vanish if a is perpendicular to both b and c.

1.11. Eigenvalues, symmetry

Properly understanding surface theory and related key concepts such as the Weingarten map (see Section 8.12) depends on linear-algebraic background related to diagonalisation of

(1) symmetric matrices or, more generally,
(2) selfadjoint endomorphisms.

The following fact was discussed in linear algebra.

PROPOSITION 1.11.1. *A real matrix may have no real eigenvector or eigenvalue.*

EXAMPLE 1.11.2. The matrix of a 90 degree rotation in the plane,

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

does not have a real eigenvector for obvious geometric reasons (no direction is preserved but rather rotated).

In this section we prove the existence of a real eigenvector (and hence, a real eigenvalue) for a real symmetric matrix and a self-adjoint endomorphism.

This fact has important ramifications in surface theory, since the various notions of curvature of a surface are defined in terms of the eigenvalues of a selfadjoint endomorphism called the Weingarten map (see Sections 8.7 and 9.10) which is a coordinate-independent formulation of the notion of a symmetric matrix. Let

$$I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

[7] Klal yad yamin

be the (n,n) identity matrix. Thus

$$I = (\delta_{ij})_{\substack{i=1,\dots,n \\ j=1,\dots,n}}$$

where δ_{ij} is the Kronecker delta.

DEFINITION 1.11.3. Let B be an (n,n)-matrix. A scalar λ is called an *eigenvalue* of B if

$$\det(B - \lambda I) = 0.$$

THEOREM 1.11.4. *If $\lambda \in \mathbb{R}$ is an eigenvalue of B, then there is a vector $v \in \mathbb{R}^n$, $v \neq 0$, such that*

$$Bv = \lambda v. \tag{1.11.1}$$

The proof was given in linear algebra and is not reproduced here.

DEFINITION 1.11.5. A nonzero vector satisfying (1.11.1) is called an *eigenvector* belonging to λ.

1.12. Euclidean inner product

DEFINITION 1.12.1. The Euclidean inner product of vectors $v, w \in \mathbb{R}^n$ is defined by

$$\langle v, w \rangle = v^1 w^1 + \cdots + v^n w^n = \sum_{i=1}^{n} v^i w^i.$$

Recall that all of our vectors are *column* vectors.

LEMMA 1.12.2. *The inner product can be expressed in terms of matrix multiplication as follows: $\langle v, w \rangle = v^t\ w$.*

Recall the following theorem from basic linear algebra.

THEOREM 1.12.3. *The transpose has the following property:*

$$(AB)^t = B^t A^t.$$

Recall the following definition (see Section 1.6).

DEFINITION 1.12.4. A square matrix A is called *symmetric* if $A^t = A$.

LEMMA 1.12.5. *Let B be a real matrix. Then the following two conditions are equivalent:*

(1) *B symmetric;*
(2) *for all $v, w \in \mathbb{R}^n$, one has $\langle Bv, w \rangle = \langle v, Bw \rangle$.*

PROOF. We have

$$\langle Bv, w\rangle = (Bv)^t w = v^t B^t w = \langle v, B^t w\rangle = \langle v, Bw\rangle,$$

proving the direction (1) $\implies$ (2). Conversely, if $v = e_i$ and $w = e_j$ then $\langle Bv, w\rangle = b_{ij}$ while $\langle v, Bw\rangle = b_{ji}$, proving the other direction (1) $\Longleftarrow$ (2). $\square$

CHAPTER 2

Eigenvalues of symmetric matrices, conic sections

2.1. Finding an eigenvector of a symmetric matrix

In this section we continue with linear-algebraic preliminaries for the theory of surfaces in Euclidean space. Eigenvalues and eigenvectors were reviewed in Section 1.11.

DEFINITION 2.1.1. The Hermitian product on $\mathbb{C}^n$ is the product

$$\langle z, w\rangle = \sum_{i=1}^{n} z^i\, \overline{w^i} \quad \forall z, w \in \mathbb{C}^n, \tag{2.1.1}$$

where $\overline{w^i}$ is the complex conjugate of w^i.

Thus the Hermitian inner product on $\mathbb{C}^n$ is linear in one variable and *skew-linear* in the other.[1]

THEOREM 2.1.2. *Every real symmetric matrix possesses a real eigenvector.*

We will give two proofs of this important theorem. The first proof is simpler, more algebraic, and passes via complexification. The second proof is more geometric.

FIRST PROOF. Let $n \geq 1$, and let $B \in \mathrm{Mat}_{n,n}(\mathbb{R})$ be an $n \times n$ real symmetric matrix. As such, it defines a linear map

$$B_{\mathbb{R}} \colon \mathbb{R}^n \to \mathbb{R}^n, \quad v \mapsto Bv$$

sending a vector $v \in \mathbb{R}^n$ to the vector Bv. We prove the theorem in five steps.

Step 1. Consider the field extension $\mathbb{R} \hookrightarrow \mathbb{C}$. Via such an inclusion, we can view B as a complex matrix $B \in \mathrm{Mat}_{n,n}(\mathbb{C})$. Then the matrix B defines a complex endomorphism

$$B_{\mathbb{C}} \colon \mathbb{C}^n \to \mathbb{C}^n$$

sending a vector $v \in \mathbb{C}^n$ to the vector Bv.

[1] The sum appearing in (2.1.1) is our convention. Some texts adopt the alternative convention $\langle z, w\rangle = \sum_{i=1}^{n} \overline{z^i} w^i$ which we will not use.

Step 2. Let $\langle\ ,\ \rangle$ be the standard Hermitian inner product in $\mathbb{C}^n$ as in (2.1.1), extending the scalar product in $\mathbb{R}^n$. Since the matrix B has real coefficients, by Lemma 1.12.5, we obtain for all $z \in \mathbb{C}^n$,

$$\langle Bz, z\rangle = \langle z, Bz\rangle \tag{2.1.2}$$

by symmetry $B = B^t$.

Step 3. The characteristic polynomial $p_B(\lambda) = \det(B - \lambda I)$ of the endomorphism $B_{\mathbb{C}}$ is a polynomial of positive degree $n > 0$. By the fundamental theorem of algebra, the polynomial p_B possesses a root $\lambda_0 \in \mathbb{C}$.

Step 4. Let $z \in \mathbb{C}^n$ be an eigenvector belonging to the eigenvalue λ_0, so that $Bz = \lambda_0 z$. Since the Hermitian inner product is skew-linear in the second variable, equality (2.1.2) gives

$$\langle \lambda_0 z, z\rangle = \langle z, \lambda_0 z\rangle = \overline{\lambda}_0\langle z, z\rangle.$$

Therefore $\lambda_0\langle z, z\rangle = \bar{\lambda}_0\langle z, z\rangle$ so that $\lambda_0 = \bar{\lambda}_0$, i.e., λ_0 is a real eigenvalue.

Step 5. Since λ_0 is a real root of p_B, there exists a real eigenvector $v \in \mathbb{R}^n$ such that $Bv = \lambda_0 v$, as required. $\square$

2.1.1. A geometric proof. The material in this subsection is optional. The second proof is somewhat longer but has the advantage of being more geometric, as well as more concrete in the construction of the eigenvector. Let $S \subseteq \mathbb{R}^n$ be the unit sphere $S = \{v \in \mathbb{R}^n \colon |v| = 1\}$. Given a symmetric matrix B, define a function $f(v) = \langle v, Bv\rangle$. We are interested in its restriction to S, i.e., $f\colon S \to \mathbb{R}$. Let v_0 be a maximum of f restricted to S. We will show that v_0 is an eigenvector of B. Let $V_0^{\perp} \subseteq \mathbb{R}^n$ be the orthogonal complement of the line spanned by v_0. Let $w \in V_0^{\perp}$. Consider the curve $v_0 + tw$, $t \geq 0$ (see also a different choice of curve in Remark 2.1.3 below). Then $\frac{d}{dt}\big|_{t=0} f(v_0 + tw) = 0$ since v_0 is a maximum and w is tangent to the sphere. Now

$$\begin{aligned}
\frac{d}{dt}\Big|_0 f(v_0 + tw) &= \frac{d}{dt}\Big|_0 \langle v_0 + tw, B(v_0 + tw)\rangle \\
&= \frac{d}{dt}\Big|_0 (\langle v_0, Bv\rangle + t\langle v_0, Bw\rangle + t\langle w, Bv_0\rangle + t^2\langle w, Bv\rangle) \\
&= \langle v_0, Bw\rangle + \langle w, Bv_0\rangle \\
&= \langle Bw, v_0\rangle + \langle Bv_0, w\rangle \\
&= \langle B^t v_0, w\rangle + \langle Bv_0, w\rangle \\
&= \langle (B^t + B)v_0, w\rangle \\
&= 2\langle Bv_0, w\rangle \quad \text{by symmetry of } B.
\end{aligned}$$

Thus $\langle Bv_0, w\rangle = 0$ for all $w \in V_0^\perp$. Hence Bv_0 is proportional to v_0 and so v_0 is an eigenvector of B.

REMARK 2.1.3. Our calculation used the curve $v_0 + tw$ which, while tangent to S at v_0 (see section 6.9), does not lie on S. Therefore one needs to use instead the curve $(\cos t)v_0 + (\sin t)w$ lying on S. Then

$$\frac{d}{dt}\Big|_{t=0}\langle(\cos t)v_0 + (\sin t)w, B((\cos t)v_0 + (\sin t)w)\rangle = \cdots = \langle(B^t + B)v_0, w\rangle$$

and one argues by symmetry as before. Alternatively, one could note that the derivative is independent of the choice of curve representing the vector and therefore the original choice of linear curve is valid, as well.

2.2. Trace of product of matrices in index notation

First we reinforce the material on index notation and Einstein summation convention defined in Section 1.3. The following result is important in its own right. We reproduce it here because its proof is a good illustration of the uses of the Einstein index notation. Recall (Definition 1.5.2) that the trace of a square matrix $A = (a^i{}_j)$ is $\operatorname{tr}(A) = a^k{}_k$. Here k is a summation index (any other letter could have been used in place of k).

THEOREM 2.2.1. *Square matrices A and B of the same size satisfy* $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.

PROOF. Let $A = (a^i{}_j)$ and $B = (b^i{}_j)$. Then

$$AB = (a^i{}_k b^k_j)_{\substack{i=1,\ldots,n\\ j=1,\ldots,n}}$$

By definition of trace (see Definition 1.5.2), we obtain

$$\operatorname{tr}(AB) = a^i{}_k b^k_i. \qquad (2.2.1)$$

Similarly,

$$\operatorname{tr}(BA) = \operatorname{tr}(b^k_i a^i{}_j) = b^k_i a^i{}_k = a^i{}_k b^k_i. \qquad (2.2.2)$$

Comparing the outcomes of calculations (2.2.1) and (2.2.2) we conclude that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$. □

EXERCISE 2.2.2. A matrix A is called *idempotent* if $A^2 = A$. Write the idempotency condition in indices with Einstein summation convention (without Σ's), keeping track of free indices and internal summation indices.

EXERCISE 2.2.3. Matrices A and B are *similar* if there exists an invertible matrix P such that $AP - PB = 0$. Write the similarity condition in indices, as the vanishing of each (i, j)th coefficient of the difference $AP - PB$.

2.3. Inner product spaces and self-adjoint endomorphisms

In Section 2.1 we worked with real matrices and showed that the symmetry of a matrix guarantees the existence of a real eigenvector. In a more general situation where a natural basis is not available, a similar statement holds for a special type of endomorphism of a real vector space with an inner product.

DEFINITION 2.3.1. Let $(V, \langle\ ,\ \rangle)$ be a real inner product space. An endomorphism $B\colon V \to V$ is called *selfadjoint* if one has

$$\langle Bv, w\rangle = \langle v, Bw\rangle \quad \forall v, w \in V. \tag{2.3.1}$$

The following is a consequence of Theorem 2.1.2.

COROLLARY 2.3.2. *Every selfadjoint endomorphism of a real inner product space admits a real eigenvector.*

PROOF. The selfadjointness was the relevant property in the proof of Theorem 2.1.2 on symmetric matrices. □

2.4. Orthogonal diagonalisation of symmetric matrices

Let $(V, \langle\ ,\ \rangle)$ be a real inner product space. Consider an endomorphism $E\colon V \to V$.

DEFINITION 2.4.1. A subspace $U \subseteq V$ is *invariant* under E if $E(U) \subseteq U$.

In other words, for every $x \in U$, one has $E(x) \in U$, as well.

DEFINITION 2.4.2. The *orthogonal complement* $O \subseteq V$ of a subspace $U \subseteq V$ is defined by $O = \{x \in V : \langle x, u\rangle = 0 \quad \forall u \in U\}$.

Such a situation is represented by the formula

$$V = U + O.$$

EXAMPLE 2.4.3. Let $U \subseteq \mathbb{R}^3$ be the (x, y)-plane. Then its orthogonal complement is the z-axis: $O = \mathbb{R}e_3$, and $\mathbb{R}^3 = U + O$.

The following theorem is known from linear algebra.

THEOREM 2.4.4. *If U and O are orthogonal complements of each other in V then* $\dim V = \dim U + \dim O$.

EXAMPLE 2.4.5 (Generalisation of Example 2.4.3). The orthogonal complement of the plane

$$U = \{(x, y, z) : ax + by + cz = 0\}$$

in $\mathbb{R}^3$ is the line O spanned by the vector $(a, b, c)^t \neq 0$.

LEMMA 2.4.6. *Let $E\colon V \to V$ be a* selfadjoint *endomorphism of an inner product space V. Suppose $U \subseteq V$ is an E-invariant subspace. Then the orthogonal complement of U in V is also E-invariant.*

PROOF. Let w be orthogonal to the subspace $U \subseteq V$. Let $u \in U$ be an arbitrary vector. Then by selfadjointness,

$$\langle E(w), u\rangle = \langle w, E(u)\rangle = 0$$

since $E(u) \in U$ by E-invariance. Therefore the vector $E(w)$ is also orthogonal to the vector u. Since this is valid for all $u \in U$, the vector $E(w)$ also belongs to the orthogonal complement of U. Hence the orthogonal complement of U is E-invariant, as required. $\square$

DEFINITION 2.4.7. A real square matrix P is *orthogonal* if $PP^t = I$, i.e., $P^{-1} = P^t$.

THEOREM 2.4.8. *Every real symmetric matrix can be orthogonally diagonalized.*

PROOF. The proof will use self-adjoint endomorphisms and invariant subspaces. A symmetric $n \times n$ matrix $S \in \mathrm{Mat}_{n,n}(\mathbb{R})$ defines a selfadjoint endomorphism $S_\mathbb{R}$ of the real inner product space $V = \mathbb{R}^n$ given by

$$S_\mathbb{R}\colon V \to V, \quad v \mapsto Sv.$$

We will give a proof in four steps.

Step 1. By Corollary 2.3.2, every selfadjoint endomorphism has a real eigenvector $v_1 \in V$, which we can assume to be a unit vector:

$$|v_1| = 1.$$

Let $\lambda_1 \in \mathbb{R}$ be its eigenvalue.

Step 2. We inductively construct a sequence of nested invariant subspaces as follows. Let $V_1 = V$ and set

$$V_2 \subseteq V_1$$

be the orthogonal complement of the line $\mathbb{R}v_1 \subseteq V_1$. Thus we have an orthogonal decomposition

$$V_1 = \mathbb{R}v_1 + V_2.$$

By Lemma 2.4.6, the subspace V_2 is invariant under the endomorphism $S_\mathbb{R}$. Consider the restriction

$$S_\mathbb{R}\restriction_{V_2}$$

of $S_\mathbb{R}$ to V_2. The restriction $S_\mathbb{R}\restriction_{V_2}$ is still selfadjoint by inheriting the property (2.3.1) restricted to V_2. Namely, since property (2.3.1) holds for all vectors $v, w \in V$, in particular it holds if these vectors are constrained to vary in a subspace $V_2 \subseteq V$. Thus we have

$$\langle S_{\mathbb{R}} v, w\rangle = \langle v, S_{\mathbb{R}} w\rangle \quad \forall v, w \in V_2. \tag{2.4.1}$$

Therefore we can apply Corollary 2.3.2 to the selfadjoint endomorphism $S_{\mathbb{R}}\restriction_{V_2}$ of V_2. As before, we find an eigenvector $v_2 \in V_2$ of $S_{\mathbb{R}}\restriction_{V_2}$ with eigenvalue $\lambda_2 \in \mathbb{R}$. Let $V_3 \subseteq V_2$ be the orthogonal complement of the line $\mathbb{R}v_2$ spanned by v_2. Next we choose an eigenvector $v_3 \in V_3$, etc. Arguing inductively, we obtain a strictly decreasing, or nested,[2] sequence of spaces

$$V_1 \supset V_2 \supset V_3 \supset \ldots \supset V_i \supset \ldots \supset \{0\}$$

which necessarily ends with a point since V_1 is finite-dimensional.

Step 3. We thus obtain an orthonormal basis consisting of unit eigenvectors $v_1, \ldots, v_n \in \mathbb{R}^n$ belonging respectively to the eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ of the endomorphism $S_{\mathbb{R}}$. Let

$$P = [v_1 \ldots v_n]$$

be the orthogonal $n \times n$ matrix whose columns are the vectors v_i, so that we have

$$P^{-1} = P^t. \tag{2.4.2}$$

Step 4. Consider the diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. By construction, we have

$$S = P\Lambda P^t$$

from (2.4.2), or equivalently,

$$SP = P\Lambda. \tag{2.4.3}$$

Indeed, to verify the relation (2.4.3), note that both sides of (2.4.3) are equal to the square matrix $[\lambda_1 v_1\ \lambda_2 v_2\ \ldots\ \lambda_n v_n]$. $\square$

2.5. Classification of conic sections: diagonalisation

We will now apply the linear-algebraic tools developed in the previous sections in order to classify conic sections in the plane up to orthogonal transformations (rotations) and translations of the plane.

Definition 2.5.1. A conic section[3] (or *conic* for short) in the plane is by definition a curve defined by the following master equation (general equation) in the (x, y)-plane:

$$ax^2 + 2bxy + cy^2 + dx + ey + f = 0, \quad a, b, c, d, e, f \in \mathbb{R}. \tag{2.5.1}$$

[2] mekunenet

[3] chatach charut

Here we chose the coefficient of the xy term to be $2b$ rather than b so as to simplify formulas like (2.5.2) below.[4]

EXAMPLE 2.5.2. We have the following examples of conic sections:
(1) a circle $x^2 + y^2 - 1 = 0$,
(2) an ellipse $x^2 + 3y^2 - 1 = 0$,
(3) a parabola $x^2 + y = 0$,
(4) a hyperbola $2xy + 1 = 0$,
(5) a hyperbola $x^2 - 5y^2 - 1 = 0$.

Consider again the master equation (2.5.1). We now let $X = \begin{pmatrix} x \\ y \end{pmatrix}$ and let

$$S = \begin{pmatrix} a & b \\ b & c \end{pmatrix}. \tag{2.5.2}$$

Then $X^t S X = ax^2 + 2bxy + cy^2$. The master equation can then be re-written as

$$X^t S X + dx + ey + f = 0. \tag{2.5.3}$$

We can now eliminate the mixed term xy as follows.

THEOREM 2.5.3. *Up to an orthogonal transformation resulting in new coordinates* (x', y'), *every conic section as in* (2.5.3) *can be written in a "diagonal" form*

$$\lambda_1 x'^{\,2} + \lambda_2 y'^{\,2} + d'x' + e'y' + f = 0, \tag{2.5.4}$$

where the coefficients λ_1 *and* λ_2 *are the eigenvalues of the matrix* S *of* (2.5.2).

PROOF. We give a proof in four steps.

Step 1. Consider the row vector

$$T = \begin{pmatrix} d & e \end{pmatrix}.$$

Then $TX = dx + ey$. Thus equation (2.5.1) becomes

$$X^t S X + TX + f = 0. \tag{2.5.5}$$

[4] One can also consider the projectivisation of the conic, namely the locus in projective coordinates of the equation $ax_1^2 + 2bx_1x_2 + cx_2^2 + dx_1x_3 + ex_2x_3 + fx_3^2 = 0$ where $a, b, c, d, e, f \in \mathbb{R}$. Here the relevant matrix is $\begin{pmatrix} a & b & \frac{d}{2} \\ b & c & \frac{e}{2} \\ \frac{d}{2} & \frac{e}{2} & f \end{pmatrix}$. If the matrix is invertible, it is either (1) definite or (2) indefinite, i.e., has both positive and negative eigenvalues. In the first case, the projective conic is empty. In the second case, it is unique up to projective transformation.

Step 2. We apply Theorem 2.4.8 to orthogonally diagonalize the symmetric matrix S to obtain $S = P\Lambda P^t$ with

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}. \tag{2.5.6}$$

Substituting this expression for S into (2.5.5) yields

$$X^t P \Lambda P^t X + TX + f = 0.$$

Step 3. We set $X' = P^t X$. Then $X = PX'$ since P is orthogonal. Furthermore, we have

$$(X')^t = (P^t X)^t = (X^t)(P^t)^t = (X^t)P.$$

Hence we obtain the equation

$$(X')^t \Lambda X' + TPX' + f = 0.$$

Step 4. Letting $T' = TP$, we obtain

$$(X')^t \Lambda X' + T'X' + f = 0,$$

where Λ is the diagonal matrix of (2.5.6). Letting x' and y' be the components of X', i.e. $X' = \begin{pmatrix} x' \\ y' \end{pmatrix}$, and $T = \begin{pmatrix} d' & e' \end{pmatrix}$, we obtain formula (2.5.4), as required. $\square$

EXAMPLE 2.5.4. The hyperbola $2xy - 1 = 0$ is not in "diagonal" form. The corresponding matrix S is $S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Here the eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = -1$. By Theorem 2.5.3 we obtain the "diagonal" equation

$$x'^2 - y'^2 - 1 = 0$$

in the new coordinates (x', y'). The result could be obtained directly by using the substitution $x = \frac{x'+y'}{\sqrt{2}}$ and $y = \frac{x'-y'}{\sqrt{2}}$.

To obtain more precise information about the conic, we need to specify certain nondegeneracy conditions, as discussed in Section 2.6.

2.6. Classification of conics: trichotomy, nondegeneracy

We apply the diagonalisation result of Section 2.5 to classify conic sections into three types (under suitable nondegeneracy conditions): ellipse, parabola, hyperbola. Such a result can be referred to as trichotomy.[5] Let S be the matrix (2.5.2).

[5]trichotomia

THEOREM 2.6.1. *If S is invertible then, up to an orthogonal transformation and a translation, the conic section can be written in the form*

$$\lambda_1(x'')^2 + \lambda_2(y'')^2 + f'' = 0 \tag{2.6.1}$$

where λ_1 and λ_2 are the eigenvalues of S.

Note that the coefficients λ_1 and λ_2 are the same as in (2.5.4) but the constant term is changed from f to f''.

PROOF. If S is invertible then both eigenvalues λ_1, λ_2 are nonzero. The term $d'x'$ in (2.5.4) can be absorbed into the quadratic term $\lambda_1 x'^2$ by completing the square as follows. We write

$$\begin{aligned}\lambda_1 x'^2 + d'x' &= \lambda_1\left(x'^2 + 2\frac{d'}{2\lambda_1}x'\right)\\ &= \lambda_1\left(x'^2 + 2\frac{d'}{2\lambda_1}x' + \left(\frac{d'}{2\lambda_1}\right)^2\right) - \lambda_1\left(\frac{d'}{2\lambda_1}\right)^2\\ &= \lambda_1\left(x' + \frac{d'}{2\lambda_1}\right)^2 - \frac{d'^2}{4\lambda_1}.\end{aligned}$$

Then we set

$$x'' = x' + \frac{d'}{2\lambda_1}.$$

Similarly $e'y'$ can be absorbed into $\lambda_2 y'^2$. Geometrically this corresponds to a translation along the axes x' and y', proving the theorem. □

DEFINITION 2.6.2. A conic section of type (2.6.1) is called a *hyperbola* if $\lambda_1\lambda_2 < 0$, provided the following nondegeneracy condition is satisfied: the constant f'' in equation (2.6.1) is nonzero.

COROLLARY 2.6.3 (Degenerate case). *Assume* $\det(S) < 0$. *If the constant f'' is zero, then instead of a hyperbola, the solution set is a degenerate conic given by a pair of transverse lines.*

REMARK 2.6.4. The transverse lines as in Corollary 2.6.3 are not necessarily orthogonal. More specifically, they are orthogonal if and only if $\lambda_1 = -\lambda_2$.

DEFINITION 2.6.5. A conic section is called an *ellipse* if $\lambda_1\lambda_2 > 0$, provided the following nondegeneracy condition is satisfied: the constant f'' in equation (2.6.1) is nonzero and has the opposite sign as compared to the sign of λ_1, i.e., $f''\lambda_1 < 0$.

See Example 12.6.1 for an application. We summarize the degenerate elliptic case as follows.

Corollary 2.6.6 (Degenerate case). *Assume* $\det(S) > 0$*.*

(1) *If* $f'' = 0$ *then instead of an ellipse one obtains a single point* $x'' = y'' = 0$*.*
(2) *If* $f'' \neq 0$ *and* f'' *has the same sign as* λ_1 *then the solution set is empty.*

We summarize our conclusions for the ellipse and the hyperbola in the following corollary. Recall that the determinant of the matrix S is the product of its eigenvalues: $\det(S) = ac - b^2 = \lambda_1\lambda_2$.

Corollary 2.6.7 (Case $\det(S) > 0$). *We have the following relations between the algebraic equation and the nature of the corresponding conic:*

(1) *If the conic* $ax^2 + 2bxy + cy^2 + dx + ey + f = 0$ *is an ellipse then* $ac - b^2 > 0$*.*
(2) *If* $ac - b^2 > 0$ *and the solution locus is neither empty nor a single point, then it is an ellipse.*

Corollary 2.6.8 (Case $\det(S) < 0$). *We have the following relations between the algebraic equation and the nature of the corresponding conic:*

(1) *If the conic* $ax^2 + 2bxy + cy^2 + dx + ey + f = 0$ *is a hyperbola then* $ac - b^2 < 0$*.*
(2) *If* $ac - b^2 < 0$ *and the solution locus is not a pair of transverse lines, then the conic is a hyperbola.*

The remaining case of vanishing determinant will be treated in Section 2.7.

2.7. Characterisation of parabolas

We continue analyzing the quadratic curve defined as the locus of the equation $ax^2 + 2bxy + cy^2 + dx + ey + f = 0$. Suppose the matrix $S = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ as in (2.5.2) satisfies $\det(S) = 0$. In this case, one cannot necessarily eliminate the linear terms by completing the square as in the cases of ellipse and hyperbola treated in Section 2.6. Therefore we continue working with the equation

$$\lambda_1 x'^2 + \lambda_2 y'^2 + d'x' + e'y' + f = 0 \tag{2.7.1}$$

from (2.5.4) where now one of the eigenvalues λ_i of S vanishes. Note that diagonalizing S does not depend on S being invertible and only requires symmetry.

DEFINITION 2.7.1. The conic section (2.7.1) is a *parabola* if the following two conditions are satisfied by the coefficients in (2.7.1):

(1) the matrix S is of rank 1 (equivalently, $\lambda_1\lambda_2 = 0$ and $\lambda_1 \neq \lambda_2$);
(2) either $\lambda_1 e' \neq 0$ or $\lambda_2 d' \neq 0$.

EXAMPLE 2.7.2. If $\lambda_1 = 1$ and $e' = 1$ we obtain the parabola $x'^2 + y' = 0$.

Parabolas are nondegenerate by definition. The following corollary describes the remaining degenerate cases.

COROLLARY 2.7.3 (Degenerate case). *Suppose S is of rank 1 and after absorbing the linear term in (2.7.1) the equation becomes $\lambda_1 x''^2 + f'' = 0$ or $\lambda_2 y''^2 + f'' = 0$. Then the solution set is either empty, a line, or a pair of parallel lines.*

EXAMPLE 2.7.4. We illustrate the three degenerate types by examples.

(1) The equation $x^2+1 = 0$ has an empty solution set in the (x, y)-plane.
(2) The equation $x^2 = 0$ has as solution set a single line, namely the y-axis.
(3) The solution set of the equation $x^2 - 1 = 0$ is a pair of parallel lines.

We mention the following result on nondegenerate quadratic curves for general culture. The result explains the term *conic section.*

THEOREM 2.7.5. *Every* nondegenerate *quadratic curve (ellipse, hyperbola, parabola) given by the solution set of the equation*

$$ax^2 + 2bxy + cy^2 + dx + ey + f = 0$$

can be represented (up to orthogonal transformation) by the intersection of a suitable plane in $\mathbb{R}^3$ and the standard cone defined by the equation $z^2 = x^2 + y^2$.

REMARK 2.7.6. Some *degenerate* cases cannot be represented by such an intersection. For example, the pair of parallel lines occurring in Corollary 2.7.3 cannot be represented as the intersection of a plane and the cone, and neither can the empty set.

CHAPTER 3

Quadric surfaces, Hessian, representation of curves

Before we go on to quadric (i.e., quadratic) surfaces, we summarize the results obtained for quadratic curves in the previous chapter.

3.1. Summary: classification of quadratic curves

The analysis of the cases presented in Sections 2.6 and 2.7 results in the following classification.

THEOREM 3.1.1. *Let $a, b, c, d, e, f \in \mathbb{R}$, and assume not all of a, b, c are* 0*. A real quadratic curve section $ax^2 + 2bxy + cy^2 + dx + ey + f = 0$ is represented by one of the following possible sets:*

(1) *the empty set $\varnothing$;*
(2) *a single point;*
(3) *union of a pair of transverse lines;*
(4) *a single line or a pair of parallel lines;*
(5) *ellipse (and then $ac - b^2 > 0$);*
(6) *parabola (and then $ac - b^2 = 0$);*
(7) *hyperbola (and then $ac - b^2 < 0$).*

REMARK 3.1.2. The first four cases are known as degenerate cases. A parabola can occur only if $ac - b^2 = 0$. An ellipse can occur only if $ac - b^2 > 0$. A hyperbola can occur only if $ac - b^2 < 0$.

3.2. Quadric surfaces

Quadric surfaces[1] are a rich source of examples that will help us illustrate basic notions of differential geometry such as Gaussian curvature and geodesic curve.

DEFINITION 3.2.1. A quadric surface $M \subseteq \mathbb{R}^3$ is the locus of points (x, y, z) satisfying the master equation

$$ax^2 + 2bxy + cy^2 + 2dxz + fz^2 + 2gyz + hx + iy + jz + k = 0, \qquad (3.2.1)$$

where $a, b, c, d, f, g, h, i, j, k \in \mathbb{R}$.

[1]mishtachim ribu'im

To bring equation (3.2.1) to standard form, we apply an orthogonal diagonalisation procedure similar to that employed in Section 2.6.

DEFINITION 3.2.2. We define matrices S, X, and T by setting

$$S = \begin{pmatrix} a & b & d \\ b & c & g \\ d & g & f \end{pmatrix}, \quad X = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad T = \begin{pmatrix} h & i & j \end{pmatrix}. \tag{3.2.2}$$

Then the quadratic part of (3.2.1) becomes $X^t SX$, and the linear part becomes TX. Then equation (3.2.1) takes the form

$$X^t SX + TX + k = 0 \tag{3.2.3}$$

as in the case of the curves.

THEOREM 3.2.3. *By means of an orthogonal transformation, the general equation* (3.2.1) *of a quadric surface M can be simplified to*

$$\lambda_1 x^2 + \lambda_2 y^2 + \lambda_3 z^2 + dx + fy + gz + k = 0, \tag{3.2.4}$$

with new variables x, y, z and new coefficients $\lambda_1, \lambda_2, \lambda_3, d, f, g, h \in \mathbb{R}$,[2] *and the same k as in* (3.2.3)*, where $\lambda_i, i = 1, 2, 3$ are the eigenvalues of S.*

PROOF. We orthogonally diagonalize the symmetric matrix S as in Section 2.5. □

3.3. Case of eigenvalues $(+++)$ or $(---)$, ellipsoid

DEFINITION 3.3.1. A quadric surface M is an *ellipsoid* if the following three conditions are satisfied:

(1) the coefficients $\lambda_i, i = 1, 2, 3$ in equation (3.2.4) are all nonzero (i.e., $\det(S) \neq 0$);
(2) all three have the same sign; and
(3) the solution locus is neither a single point nor the empty set.

An ellipsoid is by definition nondegenerate.

THEOREM 3.3.2. *Suppose* $\det S \neq 0$ *in the master equation* (3.2.3) *a quadric surface M. Then an orthogonal transformation and translation of the coordinates reduce the equation to the form*

$$\lambda_1 x^2 + \lambda_2 y^2 + \lambda_3 z^2 + \ell = 0,$$

for new variables x, y, z. If in addition the eigenvalues have the same sign, the following three cases can then occur for the quadratic surface:

[2]We do not add $'$ to the variables and constants so as to avoid encumbering the formula unnecessarily.

(1) *an ellipsoid if* $\lambda_1 \ell < 0$*;*
(2) *a degenerate surface given by a single point if* $\ell = 0$*;*
(3) *the empty set when* $\lambda_1 \ell > 0$*.*

PROOF. By Theorem 3.2.3 there exists an orthogonal transformation diagonalizing S. Next, we use the nonvanishing of the eigenvalues to complete the square as in Section 2.6, so as to eliminate the first-order term T in equation (3.2.3). □

EXAMPLE 3.3.3. We have the following examples of ellipsoids:

(1) the equation $3x^2 + 5y^2 + 7z^2 - 1 = 0$ defines an ellipsoid;
(2) equivalently $-3x^2 - 5y^2 - 7z^2 + 1 = 0$ gives the same ellipsoid as in (1);
(3) the equation $3x^2 + 5y^2 + 7z^2 = 0$ degenerates to a single point (the origin),
(4) equation $3x^2 + 5y^2 + 7z^2 + 1 = 0$ is the empty degenerate quadric surface.

3.4. Determining type of quadric surface: explicit example

To present an application of Theorem 3.3.2, let us calculate out an explicit example. Consider the surface $M \subseteq \mathbb{R}^3$ defined by the equation

$$3x^2 + y^2 - 2xz + 3z^2 - 5 = 0. \tag{3.4.1}$$

Let us determine the type of surface it is.

Step 1. The corresponding symmetric matrix $S \in \mathrm{Mat}_{3,3}(\mathbb{R})$ is

$$S = \begin{pmatrix} 3 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 3 \end{pmatrix}$$

and defines a self-adjoint endomorphism $S + \mathbb{R} \colon \mathbb{R}^3 \to \mathbb{R}^3$.

Step 2. Note that the equation is partly diagonalized already because there are no xy or yz terms. Thus the y-axis is an invariant subspace of S. Namely, the y-axis is the eigenspace of the eigenvalue $\lambda_1 = +1$, so that $Se_2 = e_2$. Its orthogonal complement, the (x, z)-plane, is also invariant under S by Lemma 2.4.6, or simply by inspection.

Step 3. To find the remaining two eigenvalues of S, we restrict the endomorphism to the (x, z) plane. Here we obtain the equation

$$3x^2 - 2xz + 3z^2 - 5 = 0. \tag{3.4.2}$$

We therefore focus on diagonalizing the quadratic part $3x^2 - 2xz + 3z^2$ of (3.4.2). Consider the corresponding symmetric matrix

$$C = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}.$$

Here we deliberately use C instead of S so as to avoid confusion with the 3×3 matrix. The 2×2 matrix C has characteristic polynomial $p_C(\lambda) = \lambda^2 - 6\lambda + 8$. Its roots are $3 \pm \sqrt{9-8} = 3 \pm 1$. Both roots $\lambda_2 = 2$ and $\lambda_3 = 4$ are positive.

Step 4. According to the general theory in dimension 2, we obtain that, after an orthogonal transformation and with respect to the new coordinates x', z', equation (3.4.2) becomes

$$2x'^2 + 4z'^2 - 5 = 0.$$

In the notation of Chapter 2, we obtain a diagonal matrix $\Lambda = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$ diagonalizing C. Note that this diagonalisation procedure does not affect the constant term, which remains unchanged. Since we are only interested in the eigenvalues, there is no need to determine x', z' explicitly (which would involve calculating the eigenvectors and finding the orthogonal matrix P).

Step 5. With respect to the new coordinates (y, x', z'), the equation of the quadric surface M of (3.4.1) takes the form

$$y^2 + 2x'^{\,2} + 4z'^{\,2} - 5 = 0.$$

Namely, we obtain a diagonal matrix $\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}$ diagonalizing S. Notice that all three eigenvalues $1, 2, 4$ of the original matrix S are positive. Furthermore, the solution set is neither a point nor the empty set since the constant term $\ell = -5$ is negative. By Theorem 3.3.2 (or Definition 3.3.1), M is an ellipsoid.

3.5. Case of eigenvalues $(-++)$ or $(--+)$, hyperboloid

Let S be a symmetric matrix as in (3.2.2) with eigenvalues λ_i, $i = 1, 2, 3$. As in Section 3.4 we assume that $\det S = \lambda_1\lambda_2\lambda_3 \neq 0$. To fix ideas, we order them in increasing order: $\lambda_1 \leq \lambda_2 \leq \lambda_3$.

Definition 3.5.1 (Degenerate quadratic surface: cone). Suppose $\lambda_1 < 0$ and $\lambda_3 > 0$ (not all of the same sign). Then

(1) The solution set of the homogeneous equation

$$\lambda_1 x^2 + \lambda_2 y^2 + \lambda_3 z^2 = 0 \tag{3.5.1}$$

is called a *cone*;

(2) A surface defined by an equation reducible to (3.5.1) by an orthogonal transformation and translation is similarly called a cone.

EXAMPLE 3.5.2. The equation $x^2 + y^2 - z^2 = 0$ defines a cone, and similarly the equation $-x^2 - y^2 + z^2 = 0$ defines (the same) cone.

DEFINITION 3.5.3. Assume for convenience that $\lambda_1 \leq \lambda_2 \leq \lambda_3$. A quadric surface M is a *hyperboloid* if $\lambda_1 < 0$ and $\lambda_3 > 0$, and M is not a cone.

A hyperboloid is nondegenerate by definition.

THEOREM 3.5.4. *Suppose* $\det(S) \neq 0$ *in the master equation of a quadric surface* M.

(1) *If all eigenvalues have the same sign and* M *is neither the empty set nor single point, then* M *is an ellipsoid;*

(2) *if the eigenvalues include both positive and negative values and* M *is not a cone, then* M *is a hyperboloid.*

PROOF. We apply orthogonal diagonalisation to S, obtaining a diagonal matrix

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

Then we complete the square so as to eliminate the linear term T as before. □

One distinguishes two types of hyperboloids as follows.

DEFINITION 3.5.5. The hyperboloid of one sheet[3] is the locus of the equation

$$az^2 = bx^2 + cy^2 - d$$

where $a > 0, b > 0, c > 0, d > 0$ (or all negative).

REMARK 3.5.6. The Gaussian curvature (see Section 9.5) of a hyperboloid of one sheet is negative at each point.

[3]chad-yeriati

DEFINITION 3.5.7. The hyperboloid of two sheets[4] is the locus of the equation

$$az^2 = bx^2 + cy^2 + d$$

where $a > 0, b > 0, c > 0, d > 0$ (or all negative).

REMARK 3.5.8. The Gaussian curvature of a hyperboloid of two sheets is positive at each point.

DEFINITION 3.5.9 (Alternative definition). Hyperboloid of k sheets, where k is 1 or 2, is the locus of the equation

$$az^2 = bx^2 + cy^2 + (-1)^k d$$

where $a > 0, b > 0, c > 0, d > 0$ (or all negative).

REMARK 3.5.10 (Sign of determinant). In this dimension (unlike the case of conic sections), the sign of the determinant $\det(S)$ has no significance and cannot be used to determine whether the surface M is an ellisoid or a hyperboloid, or which type of hyperboloid M is.

3.6. Case rank$(S) = 2$; paraboloid, hyperbolic paraboloid

Now suppose that $\operatorname{rank}(S) = 2$. This occurs if precisely two of the three eigenvalues are nonzero. We study quadric surfaces

$$X^t S X + T X + k = 0 \tag{3.6.1}$$

in the case when S has rank 2.

THEOREM 3.6.1. *Suppose matrix S in (3.6.1) has $\operatorname{rank}(S) = 2$. Up to orthogonal transformation and translation, the quadric surface M defined by $X^t S X + T X + k = 0$ takes the form*

$$\lambda_1 x^2 + \lambda_2 y^2 + cz + d = 0, \tag{3.6.2}$$

with new variables x, y, z and new coefficients c, d, where $\lambda_1 \lambda_2 \neq 0$.

PROOF. We orthogonally diagonalize the matrix as before. We relabel the new coordinates as x, y, z in such a way that

(1) the eigendirections for the nonzero eigenvalues λ_1 and λ_2 correspond to the x-axis and the y-axis, and
(2) the zero eigenvalue corresponds to the z-axis.

Thus, diagonalizing S we obtain the matrix

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

[4] du-yeriati

Next, we eliminate the linear terms in x and y by absorbing it into the respective quadratic term by completing the square as before. The linear term in z cannot be eliminated because the third eigenvalue vanishes. $\square$

DEFINITION 3.6.2. If $c \neq 0$ in equation (3.6.2), the corresponding nondegenerate quadric surface M is called a *paraboloid* (of one of two types as detailed below).

Additional special cases of quadric surfaces are the following.

REMARK 3.6.3. It is important to think through each of these examples as they will provide important illustrations of the behavior of the Gaussian curvature (to be introduced in Section 9.5) of surfaces.

DEFINITION 3.6.4. The (elliptic) *paraboloid* is the quadratic surface $z = ax^2 + by^2 + d$, where $ab > 0$.

The Gaussian curvature of the paraboloid is positive at each point.

DEFINITION 3.6.5. The *hyperbolic paraboloid* is the quadratic surface $z = ax^2 - by^2 + d$, where $ab > 0$.

The Gaussian curvature of the hyperbolic paraboloid[5] is negative at each point. See Figures 3.6.1, 3.6.2, 3.6.3.

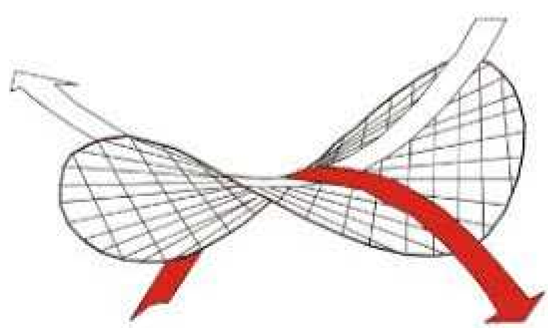

FIGURE 3.6.1. Hyperbolic paraboloid

[5]In the appendix on surfaces to the second volume of his *Introductio in analysin infinitorum* (1748), Euler gave a complete geometric classification of quadrics, and it is there that we meet the hyperbolic paraboloid for the first time. See `https://hsm.stackexchange.com/questions/15834` for details.

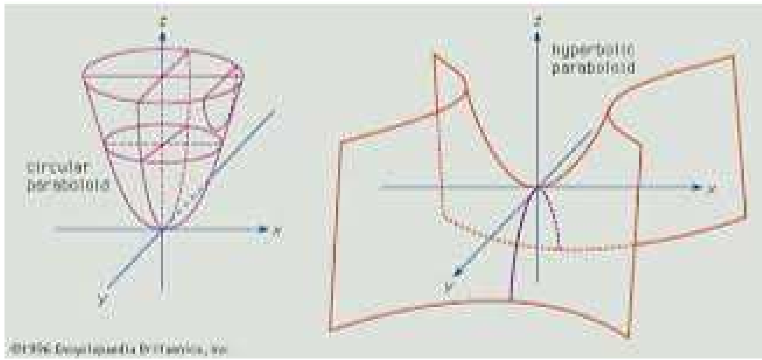

FIGURE 3.6.2. Elliptic and hyperbolic paraboloids

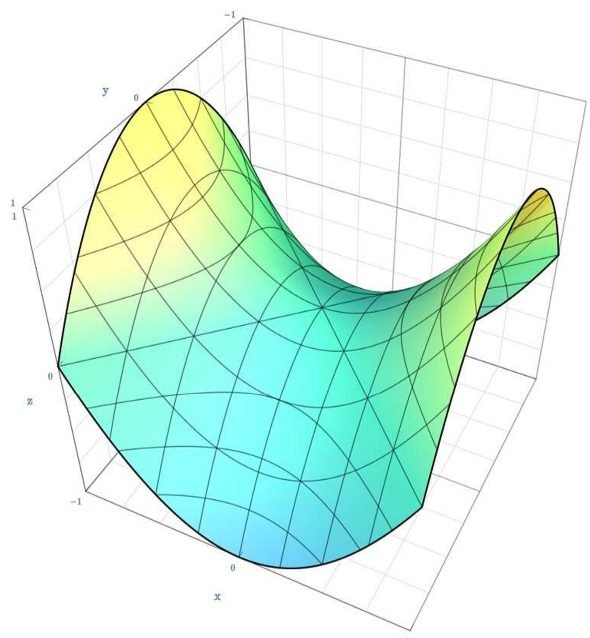

FIGURE 3.6.3. Hyperbolic paraboloid

COROLLARY 3.6.6. *Consider the surface*

$$M = \{(x, y, z) \in \mathbb{R}^3 \colon z = ax^2 + by^2\} \text{ where } ab \neq 0.$$

If $ab > 0$ then M is a (elliptic) paraboloid. If $ab < 0$ then M is a hyperbolic paraboloid.

This is immediate from the discussion above.

3.7. Cylinder

We continue studying the case $\operatorname{rank}(S) = 2$.

DEFINITION 3.7.1. Suppose $c = 0$ in equation (3.6.2), resulting in equation $\lambda_1 x^2 + \lambda_2 y^2 + d = 0$. All such surfaces in $\mathbb{R}^3$ are called *degenerate*.

EXAMPLE 3.7.2. The *cylinder* $\{(x, y, z) \in \mathbb{R}^3 \colon x^2 + y^2 - 1 = 0\}$ is an example of a degenerate quadric surface.

REMARK 3.7.3. In Section 3.1 we gave a (more-or-less) complete classification of conic sections, including the degenerate cases. For quadric surfaces, a complete classification in the case $\det(S) = 0$ is too detailed to be treated here.

3.7.1. Exercise on index notation. The material in this subsection is review.

THEOREM 3.7.4. *Every* 2×2 *matrix* $A = (a^i{}_j)$ *satisfies the identity*

$$(\forall i, j) \quad a^i{}_k a^k{}_j + \det(A)\, \delta^i{}_j = a^k{}_k a^i{}_j \tag{3.7.1}$$

Note that indices i and j and free indices, whereas k is a summation index (see Section 1.8).

PROOF. Let $p_A(\lambda)$ be the characteristic polynomial of A. For 2×2 matrices, we have $p_A(\lambda) = \lambda^2 - (\operatorname{tr} A)\lambda + \det(A)$. The Cayley-Hamilton theorem asserts that $p_A(A) = 0$. Therefore we obtain $A^2 - (\operatorname{tr} A)A + \det(A) I = 0$, which in index notation gives $a^i{}_k a^k{}_j - a^k{}_k a^i{}_j + \det A\, \delta^i{}_j = 0$. This is equivalent to (3.7.1). $\square$

EXAMPLE 3.7.5 (Exercise on index notation). For a matrix A of size 3×3, the characteristic polynomial $p_A(\lambda)$ has the form $-p_A(\lambda) = \lambda^3 - \operatorname{Tr}(A)\lambda^2 + s(A)\lambda - \det(A)\lambda^0$. Here $s(A) = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3$, where λ_i are the eigenvalues of A. By Cayley-Hamilton theorem, we have

$$p_A(A) = 0. \tag{3.7.2}$$

Express the equation (3.7.2) in index notation.

3.8. Gradient and Hessian

We review some material from vector calculus.[6] Let $(u^1, \ldots, u^n)$ be coordinates in $\mathbb{R}^n$. Consider a real-valued function $f(u^1, \ldots, u^n)$ of n variables. The function will be required to be at least C^2-smooth.

[6]This is material from infi 3, required of all math majors as it turned out on 10may'23; see 613/amirg and 613/adin.

DEFINITION 3.8.1. The *gradient* ∇f of f at a point $p = (u^1, \ldots, u^n)$ is the vector

$$\nabla f(p) = \begin{pmatrix} \frac{\partial f}{\partial u^1} \\ \frac{\partial f}{\partial u^2} \\ \vdots \\ \frac{\partial f}{\partial u^n} \end{pmatrix}$$

where each of the partial derivatives is calculated at the point p.

DEFINITION 3.8.2. A *critical point* $p \in \mathbb{R}^n$ of f is a point satisfying

$$\nabla f(p) = 0,$$

i.e. $\frac{\partial f}{\partial u^i}(p) = 0$ for all $i = 1, \ldots, n$.

We now consider the second partial derivatives of f, which we will denote $f_{ij} = \frac{\partial^2 f}{\partial u^i \partial u^j}$.

DEFINITION 3.8.3. The *Hessian matrix* H_f of f is the matrix

$$H_f = \left(f_{ij}\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}},$$

i.e., the $n \times n$ matrix of second partial derivatives.

The antisymmetrisation notation was defined in formula (1.6.2). We can then formulate what is known as Schwarz's theorem or Clairaut's theorem as follows.

THEOREM 3.8.4 (Equality of mixed partials). *Let $f \in C^2$. The following three statements are equivalent and true:*

(1) *the Hessian H_f is a symmetric matrix;*
(2) *we have $f_{ij} = f_{ji}$ for all i, j;*
(3) *in terms of the antisymmetrisation notation, $f_{[ij]} = 0$.*

3.9. Minima, maxima, saddle points

We study the graph of a function $f(u^1, u^2)$ at a critical point.

EXAMPLE 3.9.1 (maxima, minima, saddle points). Let $n = 2$. Then the *sign* of the determinant

$$\det(H_f) = \frac{\partial^2 f}{\partial u^1 \partial u^1} \frac{\partial^2 f}{\partial u^2 \partial u^2} - \left(\frac{\partial^2 f}{\partial u^1 \partial u^2}\right)^2$$

at a critical point p has geometric significance.

Namely, we have the following.

(1) If $\det(H_f(p)) > 0$ then p is a local maximum or a local minimum.

(2) If $\det(H_f(p)) < 0$, then p is a *saddle point*[7] of the graph of the function.

EXAMPLE 3.9.2 (Critical points of quadric surfaces). Quadric surfaces are a rich source of examples.

(1) The origin is a critical point for the function whose graph is the paraboloid $z = x^2 + y^2$. In the case of the paraboloid the critical point is a minimum.
(2) Similar remarks apply to the top sheet of the hyperboloid of two sheets, namely $z = \sqrt{x^2 + y^2 + 1}$, where we also get a minimum.
(3) The origin is a critical point for the function whose graph is the hyperbolic paraboloid $z = x^2 - y^2$. In the case of the hyperbolic paraboloid the critical point is a saddle point.

In addition to the sign, the *value* of $H_f(p)$ also has geometric significance in terms of an invariant we will define later called the *Gaussian curvature*, expressed by the following theorem that will be proved later.

THEOREM 3.9.3. *Let f be a function of two variables. Consider the surface $M \subseteq \mathbb{R}^3$ given by the graph $(x, y, f(x, y))$ of f in $\mathbb{R}^3$. Let $p \in \mathbb{R}^2$ be a critical point of f. Then the value of* $\det(H_f(p))$ *is precisely the Gaussian curvature of the surface M at the point $(p, f(p)) \in M$.*

See Definition 9.5.1 for more details. We will return to surfaces in Section 5.1.

3.10. Parametric representation of a curve

In Chapter 2 we studied curves given by solution sets of quadratic equations in the plane. We will now study more general curves.

REMARK 3.10.1 (Two representations). There are two main ways of representing a curve in the plane: *parametric* and *implicit*.

A curve in the plane can be represented by a pair of coordinates evolving as a function of the time parameter t:

$$\alpha(t) = (\alpha^1(t), \alpha^2(t)), \quad t \in [a, b],$$

with coordinates $\alpha^1(t)$ and $\alpha^2(t)$. Both functions are assumed to be of class C^2. Thus a parametrized curve can be viewed as a map

$$\alpha \colon [a, b] \to \mathbb{R}^2. \tag{3.10.1}$$

---

[7]Nekudat ukaf

Let C be the image of the map (3.10.1). Then $C \subseteq \mathbb{R}^2$ is the geometric curve independent of parametrisation. Thus, changing the parametrisation by setting $t = t(s)$ and replacing α by a new curve $\beta(s) = \alpha(t(s))$ preserves the geometric curve C.

Definition 3.10.2. A parametrisation $\alpha(t)$ is called *regular* if for all t one has $\alpha'(t) \neq 0$.

3.11. Implicit representation of a curve

A curve in the (x, y)-plane can also be represented implicitly as the solution set of an equation

$$F(x, y) = 0,$$

where F is a function always assumed to be of class $C^2(\mathbb{R}^2)$.

Definition 3.11.1. The level curve $C_F \subseteq \mathbb{R}^2$ is the locus of the equation

$$C_F = \{(x, y) \colon F(x, y) = 0\}.$$

Example 3.11.2. A circle of radius $r > 0$ corresponds to the choice of the function $F(x, y) = x^2 + y^2 - r^2$.

Further examples are given below.

(1) The function $F(x, y) = y - x^2$ defines a parabola.
(2) The function $F(x, y) = xy - 1$ defines a hyperbola.
(3) The function $F(x, y) = x^2 - y^2 - 1$ also defines a hyperbola.

Remark 3.11.3. In each of these cases, it is easy to find a parametrisation (at least of a part of) the level curve, by solving the equation for one of the variables. Thus, in the case of the circle, we choose the positive square root to obtain $y = \sqrt{r^2 - x^2}$, giving a parametrisation of the upper half-circle by means of the pair of formulas

$$\alpha^1(t) = t, \quad \alpha^2(t) = \sqrt{r^2 - t^2}.$$

Note this is not all of the level curve C_F.

Remark 3.11.4. Unlike the above examples, typically it is difficult to find an explicit parametrisation for a curve defined by an implicit equation $F(x, y) = 0$. Locally one can always find one in theory under a suitable nondegeneracy condition, expressed by the implicit function theorem, dealt with in Section 3.12.

3.12. Implicit function theorem

THEOREM 3.12.1 (Implicit Function Theorem). *Assume the function $F(x, y)$ is $C^1(\mathbb{R}^2)$. Suppose the gradient of F does not vanish at a specific point $p = (x, y) \in C_F$, in other words*

$$\nabla F(p) \neq 0.$$

Then there exists a regular parametrisation $(\alpha^1(t), \alpha^2(t))$ of the level curve C_F in a suitable neighborhood of p.

A useful special case is the following result.

THEOREM 3.12.2 (implicit function theorem: special case). *Assume the function $F(x, y)$ is C^1. Suppose that the partial derivative with respect to y satisfies*

$$\frac{\partial F}{\partial y}(p) \neq 0.$$

Then there exists a parametrisation $y = \alpha^2(x)$, in other words $\alpha(t) = (t, \alpha^2(t))$ of the curve C_F in a suitable neighborhood of p.

EXAMPLE 3.12.3. In the case of the circle $x^2+y^2 = r^2$, the point $(r, 0)$ on the x-axis fails to satisfy the hypothesis of Theorem 3.12.2. The curve cannot be represented by a differentiable function $y = y(x)$ in a neighborhood of this point due to the problem of a *vertical tangent*. On the other hand, a regular parametrisation still exists, e.g., $(r \cos t, r \sin t)$.

3.12.1. Jacobi's criterion. This section is optional.[8] The type of quadric surface one obtains depends critically on the signs of the eigenvalues of the matrix S. The signs of the eigenvalues can be determined without orthogonal diagonalisation, by means of Jacobi's criterion. See Giorgi [**2**, p. 57–58].

Given a matrix A over a field F, let Δ_k denote the $k \times k$ upper-left block, called a *principal minor*. Matrices A and B are *equivalent* if they are congruent (rather than similar), meaning that $B = P^t AP$ (rather than by conjugation $P^{-1}AP$). Equivalent matrices don't have the same eigenvalues (unlike similar matrices). However, the signs of the eigenvalues are preserved under equivalence. Here B is of course *not* assumed to be orthogonal.

THEOREM 3.12.4 (Jacobi). *Let $A \in \mathrm{Mat}_n(F)$ be a symmetric matrix, and assume $\det(\Delta_k) \neq 0$ for $k = 1, \ldots, n$. Then A is equivalent to the matrix diag* $\left(\frac{1}{\Delta_1}, \frac{\Delta_1}{\det\Delta_2}, \ldots, \frac{\det\Delta_{n-1}}{\det\Delta_n}\right)$.

[8]The material with regard to the Sylvester criterion (special case of Jacobi) is covered in infi 3, required of all math majors.

For example, for a 2×2 symmetric matrix $\begin{pmatrix} a & b \\ b & d \end{pmatrix}$ Jacobi's criterion affirms the equivalence to $\begin{pmatrix} \frac{1}{a} & 0 \\ 0 & \frac{a}{ad-b^2} \end{pmatrix}$.

Proof. Take the vector $b_k = \Delta_k^{-1} e_k \in F^k$ of length k, and pad it with zeros up to length n. Consider the matrix $B = (b_{ij})$ whose column vectors are $b_1, \ldots, b_n$. By Cramer's formula, the diagonal coefficients of B satisfy $b_{kk} = \det \begin{pmatrix} \Delta_{k-1} & 0 \\ 0 & 1/\det\Delta_k \end{pmatrix} = \det\Delta_{k-1}/\Delta_k$, so $\det(B) = \prod_{k=1}^n b_{kk} = 1/\det(A) \neq 0$. Compute that $B^t AB$ is lower triangular with diagonal $b_{11}, \ldots, b_{kk}$. Being symmetric, it is diagonal. $\square$

If some minor Δ_k is not invertible, then A cannot be definite. Applying this result in the case of a real symmetric matrix, we obtain the following corollary. Let A be symmetric. Then A is positive definite if and only if all $\det(\Delta_k) > 0$. Define minors in general (choose rows and columns $i_1, \ldots, i_t$). Permuting rows and columns, we obtain the following corollary.

Corollary 3.12.5. *Let A be symmetric positive definite matrix. Then all "diagonal" minors are positive definite (and in particular have positive determinants).*

An application of this is determining whether or not a quadric surface is an ellipsoid, without having to orthogonally diagonalize the matrix of coefficients, as we will see below (care has to be taken to show that the quadric surface is nondegenerate). For example, let us determine whether or not the quadric surface

$$x^2 + xy + y^2 + xz + z^2 + yz + x + y + z - 2 = 0 \tag{3.12.1}$$

is an ellipsoid. To solve the problem, we first construct the corresponding matrix $S = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix}$ and then calculate the principal minors $\Delta_1 = 1$, $\Delta_2 = 1 \cdot 1 - \frac{1}{2} \cdot \frac{1}{2} = 3/4$, and $\Delta_3 = 1 + \frac{1}{8} + \frac{1}{8} - \frac{1}{4} - \frac{1}{4} - \frac{1}{4} = \frac{1}{2}$. Thus all principal minors are positive and therefore the surface is an ellipsoid, provided we can show it is nondegenerate. To check nondegeneracy, notice that (3.12.1) has at least two distinct solutions: $(x, y, z) = (1, 0, 0)$ and $(x, y, z) = (0, 1, 0)$. Therefore it is a nondegenerate ellipsoid.

CHAPTER 4

Curvature, Laplacian, Bateman–Reiss

4.1. Unit speed parametrisation

In Section 3.10 we dealt with the notion of a parametrized curve. Let us review the notation involved.

- $\alpha\colon I \to \mathbb{R}^2$, $\alpha(t) = (\alpha^1(t), \alpha^2(t))$ is a parametrized curve in the plane;
- $C \subseteq \mathbb{R}^2$ is the underlying geometric curve, i.e., the image of α:
$$C = \left\{(x,y) \in \mathbb{R}^2 \colon (\exists t)\left(x = \alpha^1(t), y = \alpha^2(t)\right)\right\}.$$

Definition 4.1.1. We say $\alpha = \alpha(t)$ is a *unit speed* curve if $\left|\frac{d\alpha}{dt}\right| = 1$, *i.e.* $\left(\frac{d\alpha^1}{dt}\right)^2 + \left(\frac{d\alpha^2}{dt}\right)^2 = 1$ for all $t \in [a,b]$.

Definition 4.1.2. The parameter of a unit speed parametrisation, usually denoted s, is called *arclength*.

Similarly, a parametrized curve $\alpha(t)$ in $\mathbb{R}^n$ is unit speed if $|\alpha'(t)| = 1$ for all t.

Example 4.1.3 (Circle of radius r). Let $r > 0$. Then the curve
$$\alpha(s) = \left(r\cos\frac{s}{r}, r\sin\frac{s}{r}\right)$$
is a unit speed (arclength) parametrisation of the circle of radius r. Indeed, we have
$$\left|\frac{d\alpha}{ds}\right| = \sqrt{\left(r\,\frac{1}{r}\left(-\sin\frac{s}{r}\right)\right)^2 + \left(r\,\frac{1}{r}\cos\frac{s}{r}\right)^2} = \sqrt{\sin^2\frac{s}{r} + \cos^2\frac{s}{r}} = 1.$$

A regular curve always admits an arclength parametrisation; see Theorem 4.15.

4.2. Geodesic curvature of a curve

We review the standard calculus topic of the curvature of a curve.

Remark 4.2.1. The notion of the curvature of curves is closely related to the theory of surfaces in Euclidean space. It is indispensable to understanding the principal curvatures of a surface; see e.g., Theorem 9.7.1 and Theorem 9.10.5.

Our main interest will be in space curves (curves in $\mathbb{R}^3$). For such curves, one cannot in general assign a sign to the curvature. Therefore in the definition below we do not concern ourselves with the sign of the curvature of plane curves, either (see Remark 4.2.2). Signed curvature will be defined in Section 12.2.1.

In the present section, only local properties of the curvature of curves will be studied.[1]

DEFINITION 4.2.2. The (geodesic) curvature function $k_\alpha(s) \geq 0$ of a unit speed curve $\alpha(s)$ is the function

$$k_\alpha(s) = \left| \frac{d^2\alpha}{ds^2} \right|. \tag{4.2.1}$$

THEOREM 4.2.3. *The curvature $k_\alpha(s)$ of the circle of radius $r > 0$ is $k_\alpha(s) = \frac{1}{r}$ at every point of the circle.*

PROOF. With the parametrisation given in Example 4.1.3, we have

$$\frac{d^2\alpha}{ds^2} = \left(r\frac{1}{r^2}\left(-\cos\frac{s}{r} \right), r\frac{1}{r^2}\left(-\sin\frac{s}{r} \right) \right)$$

for all s, and so $k_\alpha = \left| \frac{1}{r}\left(-\cos\frac{s}{r}, -\sin\frac{s}{r} \right) \right| = \frac{1}{r}$. □

REMARK 4.2.4 (Independence of point). For the circle, the curvature is independent of the point, *i.e.* is a constant function of s.

In Section 12.3, we will give a formula for curvature with respect to an arbitrary parametrisation (not necessarily arclength).

In Section 12.4, the curvature will be expressed in terms of the angle θ formed by the tangent vector (to the curve) with the positive x-axis.

4.3. Tangent and normal vectors

Consider a plane curve $C \subseteq \mathbb{R}^2$ with arclength parametrisation $\alpha(s) = (\alpha^1(s), \alpha^2(s))$ where s is the arclength.

DEFINITION 4.3.1. The unit tangent vector $v(s)$ at a point $\alpha(s)$ of C is the vector

$$v(s) = \alpha'(s),$$

written in coordinates as $v(s) = (X(s), Y(s))$.

[1]Signed curvature: For oriented plane curves, there is a finer invariant called *signed curvature*; see Definition 12.2.1. A global result on the curvature of plane curves appears in Section 12.5.

REMARK 4.3.2. $X(s)$ and $Y(s)$ are the coordinates of $v(s)$ rather than of $\alpha(s)$. Thus $X(s) = \frac{d\alpha^1}{ds}$, $Y(s) = \frac{d\alpha^2}{ds}$.[2]

DEFINITION 4.3.3. The unit normal vector $n(s)$ to the curve $C \subseteq \mathbb{R}^2$ is the vector

$$n(s) = (Y(s), -X(s)).$$

REMARK 4.3.4 (Gauss map). Let $S^1 = \{(X, Y) \colon X^2 + Y^2 = 1\}$ be the unit circle. Since $|n(s)| = 1$ by definition, one can think of $n(s)$ as a map $C \to S^1$. This leads us to the notion of Gauss map for the curve C; see Section 12.4.

LEMMA 4.3.5. *Let $v(s) = (X(s), Y(s))$ be the unit tangent vector of the curve $\alpha(s)$, and assume $Y(s) \neq 0$. Then the following identity is satisfied:*

$$\frac{dY}{ds} = -\frac{dX}{ds}\,\frac{X}{Y}. \tag{4.3.1}$$

PROOF. We start with the formula $v(s) = (X(s), Y(s))$. Differentiating the identity $X(s)^2 + Y(s)^2 = 1$ with respect to s using chain rule, we obtain $X\frac{dX}{ds} + Y\frac{dY}{ds} = 0$ and therefore $\frac{dY}{ds} = -\frac{dX}{ds}\,\frac{X}{Y}$ as required. □

LEMMA 4.3.6. *Let $v(s) = (X(s), Y(s))$ be the unit tangent vector of $\alpha(s)$, and assume $Y(s) \neq 0$. With respect to the arclength parameter s, we have the following formula for the curvature:*

$$k_\alpha(s) = \left|\frac{1}{Y}\,\frac{dX}{ds}\right|.$$

PROOF. We use Lemma 4.3.5. Substituting the expression (4.3.1) into $\alpha'' = v'$ we obtain

$$\begin{aligned}
\alpha''(s) &= \left(\frac{dX}{ds}, \frac{dY}{ds}\right)\\
&= \left(\frac{dX}{ds}, -\frac{dX}{ds}\frac{X}{Y}\right)\\
&= \frac{dX}{ds}\left(1, -\frac{X}{Y}\right)\\
&= \frac{1}{Y}\,\frac{dX}{ds}\,(Y, -X)\\
&= \left(\frac{1}{Y}\frac{dX}{ds}\right) n
\end{aligned}$$

and therefore $k_\alpha(s) = |\alpha''(s)| = \left|\frac{1}{Y}\frac{dX}{ds}\right|$, as required. □

[2]We deliberately use capital letters X, Y here. In Section 12.4, we will use x, y for the components of the curve itself.

PROPOSITION 4.3.7. *Differentiating the normal vector*

$$n(s) = (Y(s), -X(s))$$

along the curve produces a vector proportional to $v(s)$ with factor of proportionality (up to sign) given by the geodesic curvature $k_\alpha(s)$:

$$\frac{d}{ds}n(s) = \pm k_\alpha(s)v(s). \tag{4.3.2}$$

PROOF. Differentiating and applying the formula (4.3.1) for $\frac{dY}{ds}$, we obtain

$$\begin{aligned}
-n'(s) &= \left(-\frac{dY}{ds}, \frac{dX}{ds}\right) \\
&= \left(\frac{X}{Y}\frac{dX}{ds}, \frac{dX}{ds}\right) \\
&= \frac{dX}{ds}\left(\frac{X}{Y}, 1\right) \\
&= \frac{1}{Y}\frac{dX}{ds}(X, Y) \\
&= \left(\frac{1}{Y}\frac{dX}{ds}\right)v(s),
\end{aligned}$$

and we apply Lemma 4.3.6 to prove the proposition.[3] □

REMARK 4.3.8 (Comparison to the Weingarten map for surfaces). A similar phenomenon to (4.3.2) (relating curvature to the derivative of the normal vector) occurs for a surface M. In the case of surfaces, the directional derivative of the normal vector n to M is used to define the Weingarten map (shape operator). The latter leads to the Gaussian curvature function K of M; see Section 8.12.

4.4. Osculating circle of a curve

To give a more geometric description of the curvature of a curve, we will exploit its osculating circle[4] at a point. We first recall the following fact from elementary calculus about the second derivative of a function.

THEOREM 4.4.1. *Let I be an interval, and let $f \in C^2(I)$. The second derivative f'' of a function $f(x)$ may be computed from a triple*

[3]Note that if the curve is defined implicitly by $F(x, y) = 0$, then the unit normal n to the curve satisfies $n = \frac{\nabla F}{|\nabla F|}$. There is a formula for the curvature purely in terms of F; see Section 4.8.

[4]Maagal noshek

of points $f(x), f(x+h), f(x-h)$ that are infinitely close to each other, as follows:

$$f''(x) = \lim_{h\to 0} \frac{f(x+h)+f(x-h)-2f(x)}{h^2}.$$

The theorem remains valid for vector-valued functions; see Definition 4.4.3.

Corollary 4.4.2. *The second derivative of a function can be calculated from the value of the function at a triple of nearby points x, $x+h$, $x-h$.*

Definition 4.4.3 (Osculating circle). The *osculating circle* to the curve parametrized by α with arclength parameter s at a point $p=\alpha(s)$ is obtained by choosing a circle passing through the three points

$$\alpha(s),\ \alpha(s-h),\ \alpha(s+h)$$

for infinitesimal h, i.e., taking the limit as h tends to zero.

Remark 4.4.4. The osculating circle and the curve are "better than tangent" at the point p, in the sense that they have second order tangency at p.

Corollary 4.4.5. *The curvatures of the osculating circle and the curve at the point of tangency are equal.*[5]

Proof. The curvature is defined by the second derivative. The second derivative is computed from the same triple of points for $\alpha(s)$ and for the osculating circle (see Corollary 4.4.2), proving the corollary (cf. Weatherburn [**We55**, p. 13]). $\square$

4.5. Center of curvature, radius of curvature

Recall that we denote the geometric curve by $C \subseteq \mathbb{R}^2$ and its parametrisation by $\alpha(s)$ where s is the arclength parameter.

To help geometric intuition, it is useful to recall Leibniz's and Cauchy's definition of the radius of curvature of a curve [**Cauchy 1826**]

[5]For example, let $y=f(x)$. Compute the curvature of the graph of f when $f(x)=ax^2$. Let $B=(x,x^2)$. Let A be the midpoint of OB. Let C be the intersection of the perpendicular bisector of OB with the y-axis. Let $D=(0,x)$. Triangle OAC yields $\sin\psi = \frac{OA}{OC} = \frac{\frac{1}{2}\sqrt{x^2+(ax^2)^2}}{r}$, triangle OBD yields $\sin\psi = \frac{BD}{OB} = \frac{ax^2}{\sqrt{x^2+a^2x^4}}$, and $\frac{\frac{1}{2}\sqrt{x^2+a^2x^4}}{r} = \frac{ax^2}{\sqrt{x^2+a^2x^4}}$, so that $\frac{1}{2}(x^2+a^2x^4) = arx^2$, and $\frac{1}{2}(1+a^2x^2)=ar$, so that $r=\frac{1+a^2x^2}{2a}$. Taking the limit as $x\to 0$, we obtain $r=\frac{1}{2a}$, hence $k=\frac{1}{r}=2a=f''(0)$. Thus the curvature of the parabola at its vertex equals the second derivative with respect to x (even though x is not the arclength parameter of the graph).

as the distance from the curve to the intersection point of two infinitely close normals to the curve. In more detail, we have the following.

DEFINITION 4.5.1 (Center of curvature). Let $p \in C \subseteq \mathbb{R}^2$. The *center of curvature* q of C associated with p is the intersection point of the normals to the curve at infinitely close points p and p' of C.

EXAMPLE 4.5.2 (Curvature of circle). If C is a circle, then the intersection point of two normals is always the center of C even if the normals are not close to each other. Thus the center of C and the center of curvature q (at any point) coincide. For a general curves we need to take the normals close to each other.

THEOREM 4.5.3 (Relation between center of curvature and osculating circle). *At a point $p \in C$, consider the osculating circle to the curve. Then the corresponding center of curvature q of C (the intersection point of two infinitely close normals) is the center of the osculating circle.*

DEFINITION 4.5.4 (Radius of curvature). Let $p \in C$. The *radius of curvature* R of a curve C at p is the distance from p to the center of curvature q. Namely, $R = |pq|$.

COROLLARY 4.5.5. *The curvature k_α at the point p is the inverse of the radius of curvature: $k_\alpha = \frac{1}{R} = \frac{1}{|pq|}$.*

4.6. Flat Laplacian

Computing the curvature of implicitly defined curves $C \subseteq \mathbb{R}^2$ will involve a certain second-order differential operator. We start by reviewing the simplest example of a second-order operator, the Laplacian.

DEFINITION 4.6.1. The *flat Laplacian* Δ_0 on $\mathbb{R}^2$ is the differential operator is defined in the following two equivalent ways:

(1) $\Delta_0 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$;
(2) we apply Δ_0 to a smooth function $F = F(x, y)$ by setting $\Delta_0 F = \frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2}$.

We also introduce the traditional briefer notation

$$F_{xx} = \frac{\partial^2 F}{\partial x^2}, \quad F_{yy} = \frac{\partial^2 F}{\partial y^2}, \quad F_{xy} = \frac{\partial^2 F}{\partial x\, \partial y}.$$

Then we obtain the equivalent formula

$$\Delta_0(F) = F_{xx} + F_{yy}.$$

EXAMPLE 4.6.2. If $F(x, y) = x^2 + y^2 - r^2$, then $\frac{\partial^2 F}{\partial x^2} = 2$ and similarly $\frac{\partial^2 F}{\partial y^2} = 2$, hence $\Delta_0 F = 4$.

4.7. Bateman–Reiss operator

Recall that we have two types of presentations of a curve, either parametric or implicit.

REMARK 4.7.1. In Section 4.2 the curvature of a curve was calculated starting with a parametric representation of the curve. If a curve is given implicitly as the locus (solution set) of an equation $F(x,y) = 0$, one can also calculate the geodesic curvature, by means of the theorems given in the Section 4.8, and exploiting the Bateman–Reiss operator of Definition 4.7.2.

We will be interested in the following second-order operator that appears in a formula for curvature.

DEFINITION 4.7.2. Let $F(x,y)$ be a function of two variables. The Bateman–Reiss operator D_B is defined by

$$D_B(F) = F_{xx}F_y^2 - 2F_{xy}F_xF_y + F_{yy}F_x^2. \tag{4.7.1}$$

REMARK 4.7.3. D_B is a non-linear second order differential operator.

The subscript "B" stands for Bateman, as in the Bateman equation $F_{xx}F_y^2 - 2F_{xy}F_xF_y + F_{yy}F_x^2 = 0$.

THEOREM 4.7.4. *The Bateman operator can be represented by the determinant*

$$D_B(F) = -\det \begin{pmatrix} 0 & F_x & F_y \\ F_x & F_{xx} & F_{xy} \\ F_y & F_{xy} & F_{yy} \end{pmatrix}. \tag{4.7.2}$$

PROOF. Expanding the determinant (4.7.2) along the first row, we obtain the formula (4.7.1). □

REMARK 4.7.5 (Historical remark). This operator was treated in detail in [**Goldman 2005**, p. 637, formula (3.1)]. The same operator occurs in the Reiss[6] relation in algebraic geometry; see Griffiths and Harris [**GriH78**, p. 677]. See also Valenti [**7**, p. 804] who refers to geodesic curvature as *isophote curvature* in the context of a study of luminosity and eye center location. See also [**8**] and (two) references therein.

---

[6]Michel Reiss (1805–1869).

4.8. Geodesic curvature for an implicit curve

The operator D_B defined in Section 4.7 allows us to calculate the curvature of a curve presented in implicit form, without having to specify a parametrisation. Let $C_F \subseteq \mathbb{R}^2$ be a curve defined implicitly by $F(x,y) = 0$.

THEOREM 4.8.1. *Let $p \in C_F$, and suppose $\nabla F(p) \neq 0$. Then the geodesic curvature k_C of C_F at the point p is given by*

$$k_C = \frac{|D_B(F)|}{|\nabla F|^3},$$

where D_B is the Bateman–Reiss operator defined in (4.7.1).

A proof can be found in [**Goldman 2005**, p. 637].

4.9. Curvature of circle via D_B

We will use the formula for curvature based on the Bateman–Reiss operator to calculate the curvature and solve maximization problems. We start with the circle.

EXAMPLE 4.9.1. The circle C of radius r is defined by the equation $F(x,y) = 0$, where $F = x^2 + y^2 - r^2$. Then

$$F_x = 2x, \quad F_y = 2y, \quad \nabla F = (2x, 2y)^t,$$

and therefore $|\nabla F| = 2\sqrt{x^2+y^2} = 2r$. Meanwhile, $F_{xx} = 2$, $F_{yy} = 2$, $F_{xy} = 0$, hence

$$D_B(F) = 2(2y)^2 + 2(2x)^2 = 8r^2,$$

and therefore the curvature is $k_C = \frac{8r^2}{8r^3} = \frac{1}{r}$, in agreement with Theorem 4.2.3.

4.10. Curvature of parabola via D_B

Consider the parabola $C = C_F$ given by the zero locus of $F(x,y) = y - x^2$. Then

$$F_x = -2x, \quad F_y = 1, \quad |\nabla F| = \sqrt{1+4x^2}.$$

Meanwhile, $F_{xx} = -2$, $F_{yy} = 0$, $F_{xy} = 0$. Applying the formula $D_B(F) = F_{xx}F_y^2 - 2F_{xy}F_xF_y + F_{yy}F_x^2$ we obtain $D_B(F) = -2(1)^2 = -2$. We obtain the following theorem.

THEOREM 4.10.1. *The curvature of the parabola $y = x^2$ satisfies*

$$k_C = \frac{2}{(1+x^2)^{3/2}}. \tag{4.10.1}$$

This enables us to identify the point of maximal curvature in Theorem 4.10.2 as follows.

THEOREM 4.10.2. *The apex of the parabola $y = x^2$ is its point of maximal curvature.*

PROOF. By formula (4.10.1), $k_C = \frac{2}{(1+x^2)^{3/2}}$. Since $1 + x^2 \geq 1$, the curvature attains its maximal value $k = 2$ when $x = 0$. $\square$

COROLLARY 4.10.3. *The circle of curvature of the parabola at the origin has radius $R = \frac{1}{2}$.*

4.11. Curvature of hyperbola via D_B

Consider the hyperbola $C = C_F$ given by the zero locus of $F(x, y) = xy - 1$. Then

$$F_x = y, \quad F_y = x, \quad |\nabla F| = \sqrt{x^2 + y^2}.$$

Meanwhile $F_{xx} = 0$, $F_{yy} = 0$, and $F_{xy} = 1$. Hence $D_B(F) = -2xy$. To simplify the expression for $D_B(F)$ we exploit the defining equation of the curve $xy = 1$. Hence we have $D_B(F) = -2$ at every point of the hyperbola. We obtain the following theorem.

THEOREM 4.11.1. *The curvature of the hyperbola $y = \frac{1}{x}$ is*

$$k_C = \frac{2}{(x^2 + y^2)^{3/2}}. \tag{4.11.1}$$

This enables us to find the maximum in Theorem 4.11.2 as follows.

THEOREM 4.11.2. *The curvature of the hyperbola C defined by the equation $xy = 1$ is maximal at the points $(1, 1)$ and $(-1, -1)$.*

PROOF. Formula (4.11.1) gives $k_C = \frac{2}{(x^2+y^2)^{3/2}}$.

Step 1. In order to *maximize* this expression along the curve, it suffices to *minimize* the expression $x^2 + y^2$ whose power appears in the denominator, restricted to the curve itself. We can assume without loss of generality that $x > 0$. We exploit the defining relation $xy = 1$ of the curve to obtain

$$x^2 + y^2 = (x + y)^2 - 2xy = (x + y)^2 - 2 = \left(x + \tfrac{1}{x}\right)^2 - 2. \tag{4.11.2}$$

Step 2. To minimize the expression (4.11.2), it suffices to minimize the quantity $x + \frac{1}{x}$. We will show that the sum $x + \frac{1}{x}$ is minimal when $x = 1$. Namely, to show that $x + \frac{1}{x} \geq 2$, rewrite the inequality as $x^2 + 1 \geq 2x$ or $x^2 + 1 - 2x \geq 0$ or equivalently $(x - 1)^2 \geq 0$ which is a true inequality. Hence the maximum of the curvature is when $x = 1$ and so $y = 1$, proving the theorem.

The same result can be obtained by differentiating the expression $\left(x + \frac{1}{x}\right)^2 - 1$. $\square$

4.12. Maximal curvature of a logarithmic curve

THEOREM 4.12.1. *Consider the logarithmic curve*

$$C = \{(x, y) \colon y = \ln x,\ x > 0\}.$$

Then the maximum of curvature of the curve is attained at the point $x = \frac{1}{\sqrt{2}}$, $y = -\frac{\ln 2}{2}$.

PROOF. Let $F(x, y) = y - \ln x$. Then

$$F_x = -\frac{1}{x}, \quad F_y = 1, \quad |\nabla F| = \sqrt{1 + x^{-2}}.$$

Then $F_{xx} = \frac{1}{x^2}$, $F_{yy} = 0$, and $F_{xy} = 0$. Applying the formula $D_B(F) = F_{xx}F_y^2 - 2F_{xy}F_xF_y + F_{yy}F_x^2$ we obtain $D_B(F) = \frac{1}{x^2} \cdot 1 - 2 \cdot 0 + 0 = \frac{1}{x^2} = x^{-2}$. The curvature of the logarithmic curve is therefore

$$\begin{aligned} k_C &= \frac{x^{-2}}{(\sqrt{1 + x^{-2}})^3} \\ &= \frac{1}{x^2(1 + x^{-2})^{3/2}} = \frac{x^3}{x^2(x^2 + 1)^{3/2}} = \frac{x}{(x^2 + 1)^{3/2}}. \end{aligned}$$

To maximize k_C it is sufficient to maximize the expression $k_C^2 = \frac{x^2}{(x^2+1)^3}$.

We will use an auxiliary variable $z = x^2$ to simplify calculations. We are therefore interested in maximizing the function

$$g(z) = \frac{z}{(z + 1)^3}.$$

Differentiating we obtain

$$\begin{aligned} g'(z) &= \frac{(z + 1)^3 \cdot 1 - z \cdot 3(z + 1)^2}{(z + 1)^6} \\ &= \frac{z + 1 - 3z}{(z + 1)^4} = \frac{1 - 2z}{(z + 1)^4} = 0. \end{aligned}$$

We obtain an extremum when $1 - 2z = 0$, i.e., $z = \frac{1}{2}$. Checking that the second derivative is negative at the point, we conclude that this is a point of maximum. Thus the maximum of the curvature of the logarithmic curve is attained when $x^2 = \frac{1}{2}$, i.e., $x = 2^{-1/2}$. Hence $y = \ln(2^{-1/2}) = -\frac{1}{2}\ln 2$. $\square$

4.13. Exploiting defining equation in studying curvature

The defining equation of the curve may be exploited several times in the course of the calculation when studying the curvature of a curve. Let us calculate out a specific example as follows.

THEOREM 4.13.1. *Let $a > 0$ be fixed. The maxima of the curvature of the curve*

$$C = \{(x, y) \colon xy + y^2 - a = 0\} \tag{4.13.1}$$

are at the two points such that $y = \pm \frac{\sqrt{a}}{\sqrt[4]{2}}$ while the corresponding x can be found from the defining equation.

PROOF. Set $F(x, y) = xy + y^2 - a$. Then

$$F_x = y,\ F_y = x{+}2y,\ |\nabla F| = \sqrt{y^2 + (x + 2y)^2} = \sqrt{y^2 + x^2 + 4xy + 4y^2}$$

and therefore

$$|\nabla F| = \sqrt{y^2 + x^2 + 4(xy + y^2)}. \tag{4.13.2}$$

We will give a proof in four steps.

Step 1. We use the defining equation of the curve to replace $xy+y^2$ by a. This enables us to simplify (4.13.2) to

$$|\nabla F| = \sqrt{x^2 + y^2 + 4a}. \tag{4.13.3}$$

Next, $F_{xx} = 0$, $F_{xy} = 1$, and $F_{yy} = 2$. Therefore

$$D_B(F) = 0 - 2xy - 4y^2 + 2y^2 = -2(xy + y^2). \tag{4.13.4}$$

Step 2. Exploiting the defining relation of the curve, we obtain from equation (4.13.4) that

$$D_B(F) = -2a. \tag{4.13.5}$$

Using (4.13.3) and (4.13.5), we obtain that the curvature at a point (x, y) of the curve is

$$k_C = \frac{|D_B(F)|}{|\nabla F|^3} = \frac{2a}{(x^2 + y^2 + 4a)^{3/2}}. \tag{4.13.6}$$

The curvature (4.13.6) of the curve is maximal when the expression $x^2 + y^2 + 4a$ along the curve is minimal, or equivalently when $x^2 + y^2$ is minimal along the curve. At this point one can either apply Lagrange multipliers[7] or simply solve for x, as in Step 3 below.

[7]Lagrange multipliers were treated in Infi 3 (88-230), required of all math majors.

Step 3. Using the defining relation of the curve we obtain $xy = a - y^2$ or

$$x = \frac{a - y^2}{y} = \frac{a}{y} - y. \tag{4.13.7}$$

Therefore along the curve, we obtain $x^2+y^2 = \frac{(a-y^2)^2}{y^2}+y^2 = \frac{(a-y^2)^2+y^4}{y^2}$. To simplify calculations, we introduce an auxiliary variable $z = y^2$. Thus we have

$$x^2 + y^2 = \frac{(a-z)^2 + z^2}{z}. \tag{4.13.8}$$

Step 4. The expression (4.13.8) to be minimized is

$$g(z) = \frac{(a-z)^2 + z^2}{z} = \frac{2z^2 - 2az + a^2}{z} = 2z + \frac{a^2}{z} - 2a.$$

We set $g'(z) = 2 - \frac{a^2}{z^2} = 0$ to find the extremum $z = \frac{a}{\sqrt{2}}$. One checks that the second derivative is positive. Since $z = y^2$, we obtain that $y = \pm\sqrt{z} = \pm\frac{\sqrt{a}}{2^{1/4}}$. The corresponding x can be found from $x = \frac{a-y^2}{y}$. This gives the maximum of the curvature for the curve (4.13.1). $\square$

Remark 4.13.2. Another example (cusp) is calculated out in Theorem 5.11.1.

4.14. Curvature of graph of function

Theorem 4.14.1. *Let $c \in \mathbb{R}$ be a critical point of $f(x)$. Consider the graph of f at the point $(c, f(c)) \in \mathbb{R}^2$. Then the curvature k of the graph at this point equals $k = |f''(c)|$.*

Remark 4.14.2. We parametrize the graph by $\alpha(t) = (t, f(t))$. Then we have $\frac{\partial^2\alpha}{\partial t^2} = \left(0, \frac{d^2f}{dt^2}\right)$. Therefore at the critical point c, we have $\left|\frac{\partial^2\alpha}{\partial t^2}(c)\right| = \left|\frac{d^2f}{dt^2}(c)\right|$. This is the expected answer at a critical point. However, this argument is merely a heuristic calculation because the parametrisation $(t, f(t))$ is not a unit speed parametrisation of the graph.

Proof. We apply the characterisation of curvature in terms of the Bateman operator $D_B(F) = F_{xx}F_y^2 - 2F_{xy}F_xF_y + F_{yy}F_x^2$. Let $F(x,y) = -f(x) + y$. At the critical point c, we have

$$\nabla F = (-f'(c), 1)^t = (0, 1)^t,$$

while

$$D_B(F) = -f''(x)(1)^2 - 0 + 0 = -f''(x).$$

Hence the curvature of the graph at this point satisfies

$$k = \frac{|D_B F|}{|\nabla F|^3} = \frac{|f''(c)|}{1} = |f''(c)|,$$

as required. □

EXAMPLE 4.14.3. The apex of the parabola $y = x^2$ is a critical point. At the apex ($x = 0$), the second derivative is 2 and therefore the curvature is 2, while the radius of the osculating circle is $\frac{1}{2}$, which agrees with Corollary 4.10.3.

4.15. Existence of arclength

Consider a geometric curve $C \subseteq \mathbb{R}^2$. Assume C admits a regular parametrisation. We will now prove the existence of a unit speed parameter s, i.e., the arclength parameter.

LEMMA 4.15.1. *The length L of the plane curve defined by the graph of a function $f(x)$ from a to b is*

$$L = \int_a^b \sqrt{1 + f'(x)^2}\, dx\ .$$

This is proved in calculus. More generally,

LEMMA 4.15.2. *For a curve $\alpha(t) = (\alpha^1(t), \alpha^2(t))$, we have the following formula for the length:*

$$\begin{aligned} L &= \int_a^b \sqrt{\left(\frac{d\alpha^1}{dt}\right)^2 + \left(\frac{d\alpha^2}{dt}\right)^2} dt \\ &= \int_a^b \left|\frac{d\alpha}{dt}\right| dt. \end{aligned}$$

This follows from the Pythagorean theorem applied the curve.

THEOREM 4.15.3. *Let $\alpha(t)$ be a regular parametrisation of the geometric curve $C \subseteq \mathbb{R}^2$. Then there exists a unit speed parametrition $\beta(s) = \alpha(t(s))$ of the curve C, where the parameter s is defined by* $s(t) = \int_a^t \left|\frac{d\alpha}{d\tau}\right| d\tau$.

PROOF. We define the new parameter $s = s(t)$ by setting

$$s(t) = \int_a^t \left|\frac{d\alpha}{d\tau}\right| d\tau, \tag{4.15.1}$$

where τ is a dummy variable (internal variable of integration). The Fundamental Theorem of Calculus applied to (4.15.1) yields

$$\frac{ds}{dt} = \left|\frac{d\alpha}{dt}\right|.$$

The function being monotone increasing, there exists an inverse function $t = t(s)$. Let $\beta(s) = \alpha(t(s))$. Then by chain rule

$$\frac{d\beta}{ds} = \frac{d\alpha}{dt}\frac{dt}{ds} = \frac{d\alpha}{dt}\frac{1}{ds/dt} = \frac{d\alpha}{dt}\frac{1}{|d\alpha/dt|}.$$

Thus $\left|\frac{d\beta}{ds}\right| = 1$. □

We provide an example of non-existence of a regular parametrisation for a curve with a smooth parametrisation which is not regular.

Example 4.15.4 (Cusp curve). The plane curve $\alpha(t) = (t^3, t^2)$ is smooth but not regular. Its graph exhibits a cusp.[8] In this case it is impossible to find a smooth arclength parametrisation of the curve. In Theorem 5.11.1 it is shown that its curvature tends to infinity as it approaches the cusp.

Remark 4.15.5 (Curves in $\mathbb{R}^3$). A space curve may be written in coordinates as

$$\alpha(s) = (\alpha^1(s), \alpha^2(s), \alpha^3(s)).$$

Here s is the arc length if $\left|\frac{d\alpha}{ds}\right| = 1$ *i.e.* $\sum_{i=1}^3 \left(\frac{d\alpha^i}{ds}\right)^2 = 1$.

Example 4.15.6. Helix $\alpha(t) = (a\cos\omega t, a\sin\omega t, bt)$.
(i) make a drawing in case $a = b = \omega = 1$.
(ii) parametrize by arc length.
(iii) compute the curvature.

[8] chod

CHAPTER 5

Surfaces and their curvature

5.1. Surfaces; Arnold's observation on folding a page

The differential geometry of surfaces in Euclidean 3-space starts with the observation that they inherit a metric structure from the ambient space (i.e. the Euclidean space).

QUESTION 5.1.1. We would like to understand which geometric properties of this structure are intrinsic?[1]

Part of the job is to clarify the sense of the term intrinsic.

REMARK 5.1.2. Following Arnold [**Ar74**, Appendix 1, p. 301], note that a piece of paper may be placed flat on a table, or it may be rolled into a cylinder, or it may be rolled into a cone. In mathematical terms this can be formulated by saying that

> the plane, the cylinder, and the cone (apart from the vertex) have the same local intrinsic geometry.

However, the piece of paper cannot be transformed into the surface of a sphere, that is, without tearing or stretching. Understanding this phenomenon quantitatively is our goal, *cf.* Figure 12.12.1.

5.2. Regular surface; Jacobian

Consider a surface $M \subseteq \mathbb{R}^3$ parametrized by a map $\underline{x}(u^1, u^2)$ or

$$\underline{x} : \mathbb{R}^2 \to \mathbb{R}^3. \tag{5.2.1}$$

We will always assume that $\underline{x}$ is differentiable.

DEFINITION 5.2.1. The *Jacobian matrix* $J_{\underline{x}}$ is the 3×2 matrix

$$J_{\underline{x}} = \left(\frac{\partial x^i}{\partial u^j} \right), \tag{5.2.2}$$

where x^i, $i = 1, 2, 3$, are the three components of the vector valued function $\underline{x}$.

[1] This is atzmit rather than pnimit according to Vishne.

DEFINITION 5.2.2. The parametrisation $\underline{x}$ of the surface M is called *regular* if one of the following equivalent conditions is satisfied:

(1) the vectors $\frac{\partial \underline{x}}{\partial u^1}$ and $\frac{\partial \underline{x}}{\partial u^2}$ are linearly independent;
(2) the vector product $x_1 \times x_2$ is nonzero, where $x_i = \frac{\partial \underline{x}}{\partial u^i}$;
(3) the Jacobian matrix (5.2.2) is of rank 2.

EXAMPLE 5.2.3 (Regular parametrisation of hemisphere). Let $b > 0$ be a fixed real number, and let $f(x,y) = \sqrt{b^2 - x^2 - y^2}$. The graph of the function f in $\mathbb{R}^3$ can be parametrized as follows:

$$\underline{x}^+(u^1, u^2) = (u^1, u^2, f(u^1, u^2)). \tag{5.2.3}$$

This provides a parametrisation of the (open) northern hemisphere. Then

$$\begin{cases} \frac{\partial \underline{x}^+}{\partial u^1} = (1, 0, f_x(u^1, u^2))^t \\ \frac{\partial \underline{x}^+}{\partial u^2} = (0, 1, f_y(u^1, u^2))^t. \end{cases}$$

These vectors are linearly independent and therefore (5.2.3) is a regular parametrisation.

Note that we have the relations

$$f_x = \frac{-x}{f} = \frac{-x}{\sqrt{b^2 - x^2 - y^2}}, \qquad f_y = \frac{-y}{f}. \tag{5.2.4}$$

The southern hemisphere can be similarly parametrized by

$$\underline{x}^- = (u^1, u^2, -f(u^1, u^2)).$$

The formulas for the $\frac{\partial \underline{x}^-}{\partial u^i}$ are then $\frac{\partial \underline{x}^-}{\partial u^1} = (1, 0, -f_x)^t = (1, 0, \frac{x}{\sqrt{b^2-x^2-y^2}})^t$, while $\frac{\partial \underline{x}^-}{\partial u^2} = (0, 1, -f_y)^t = (0, 1, \frac{y}{\sqrt{b^2-x^2-y^2}})^t$.

5.3. Coefficients of first fundamental form of a surface

Our starting point in the analysis of the intrinsic geometry of surfaces is the first fundamental form of a surface. The first fundamental form is obtained by restricting the 3-dimensional inner product.

REMARK 5.3.1 (Measuring length of curves). What does the first fundamental form measure? A helpful observation to keep in mind is that it enables one to measure the length of curves on the surface.

Let $\langle\,,\,\rangle$ denote the inner product in $\mathbb{R}^3$. Consider a regular parametrized surface $\underline{x}(u^1, u^2)$.

DEFINITION 5.3.2. For $i = 1, 2$ and $j = 1, 2$, define functions $g_{ij} = g_{ij}(u^1, u^2)$ called "metric coefficients" of the surface by

$$g_{ij}(u^1, u^2) = \left\langle \frac{\partial \underline{x}}{\partial u^i}, \frac{\partial \underline{x}}{\partial u^j} \right\rangle. \tag{5.3.1}$$

Remark 5.3.3. We have $g_{ij} = g_{ji}$ as the inner product is symmetric.

Example 5.3.4 (Metric coefficients for the graph of a function). The surface defined by the graph of a function $f = f(x, y)$ as in Example 5.2.3 satisfies

$$\left\langle \frac{\partial \underline{x}}{\partial u^1}, \frac{\partial \underline{x}}{\partial u^1} \right\rangle = 1 + f_x^2,$$

$$\left\langle \frac{\partial \underline{x}}{\partial u^1}, \frac{\partial \underline{x}}{\partial u^2} \right\rangle = f_x f_y,$$

$$\left\langle \frac{\partial \underline{x}}{\partial u^2}, \frac{\partial \underline{x}}{\partial u^2} \right\rangle = 1 + f_y^2.$$

Therefore in this case we have the matrix of metric coefficients

$$(g_{ij}) = \begin{pmatrix} 1 + f_x^2 & f_x f_y \\ f_x f_y & 1 + f_y^2 \end{pmatrix}$$

In the case of the hemisphere, the partial derivatives are as in formula (5.2.4).

5.4. Metric coefficients in spherical coordinates

Spherical coordinates[2] (ρ, θ, φ) in 3-space will be reviewed in more detail in Section 6.8. Briefly, φ is the angle formed by the position vector of the point with the positive direction of the z-axis. Meanwhile, θ is the polar coordinate angle θ for the projection of the vector to the x, y plane.

Example 5.4.1 (Metric coefficients in spherical coordinates). Consider the parametrisation of the unit sphere S^2 by means of spherical coordinates, so that $\underline{x} = \underline{x}(\theta, \varphi)$. We have the Cartesian coordinates

$$x = \sin\varphi\cos\theta, \quad y = \sin\varphi\sin\theta, \quad z = \cos\varphi\,.$$

Setting $u^1 = \theta$ and $u^2 = \varphi$, we obtain

$$\frac{\partial \underline{x}}{\partial u^1} = (-\sin\varphi\sin\theta, \sin\varphi\cos\theta, 0)$$

and

$$\frac{\partial \underline{x}}{\partial u^2} = (\cos\varphi\cos\theta, \cos\varphi\sin\theta, -\sin\varphi),$$

Thus in this case we have

$$(g_{ij}) = \begin{pmatrix} \sin^2\varphi & 0 \\ 0 & 1 \end{pmatrix}$$

[2] Koordinatot kaduriot

so that $\det(g_{ij}) = \sin^2\varphi$ and $\sqrt{\det(g_{ij})} = \sin\varphi$. This expression will appear in the formula for the area in Section 5.

5.5. Matrix (g_{ij}) as Gram matrix

Consider a regular parametrisation $\underline{x}(u^1,u^2)$ of a surface $M \subseteq \mathbb{R}^3$. Let $p = \underline{x}(u^1,u^2) \in \mathbb{R}^3$ be a particular point on the surface.

Definition 5.5.1. The vectors $x_i = \frac{\partial \underline{x}}{\partial u^i}$, $i = 1,2$, are called the *tangent vectors* to the surface M at the point $p = \underline{x}(u^1,u^2)$.

Lemma 5.5.2. *The Jacobian matrix of* $\underline{x}$ *is* $J_{\underline{x}} = \begin{pmatrix} x_1 & x_2 \end{pmatrix}$.

This is immediate from the definition. Next, we express the relationship between the metric coefficients and the Gram matrix of the tangent vectors as follows.

Definition 5.5.3. Given an ordered n-tuple $S = (v_i)_{i=1,\ldots,n}$ in $\mathbb{R}^b$, we define its *Gram matrix* as the matrix of inner products

$$Gram(S) = (\langle v_i, v_j\rangle)_{i=1,\ldots,n;\, j=1,\ldots,n}. \tag{5.5.1}$$

In Section 5.4 we defined the metric coefficients $g_{ij} = g_{ij}(u^1,u^2)$. We now state a relationship between the Gram matrix and the metric coefficients.

Theorem 5.5.4. *The* 2×2 *matrix* (g_{ij}) *is the Gram matrix of the pair of tangent vectors:*

$$(g_{ij}) = Gram(x_1,x_2) = J_{\underline{x}}^T\ J_{\underline{x}},$$

where $J_{\underline{x}}$ *is the Jacobi matrix.*[3]

The proof is immediate.

5.6. First fundamental form as bilinear form

Definition 5.6.1. The *tangent plane* T_pM to the surface M at the point $p = \underline{x}(u^1,u^2)$ is the plane in $\mathbb{R}^3$ spanned by the vectors $x_1(u^1,u^2)$ and $x_2(u^1,u^2)$.

Definition 5.6.2. The *first fundamental form* I_p of the surface M at the point p is the bilinear form on the tangent plane T_p, namely

$$I_p : T_pM \times T_pM \to \mathbb{R},$$

defined by the restriction of the ambient Euclidean inner product:

$$I_p(v,w) = \langle v,w\rangle_{\mathbb{R}^3},$$

for all $v,w \in T_pM$.

[3] *cf.* formula (14.6.1).

LEMMA 5.6.3. *With respect to the basis (x_1, x_2), the first fundamental form is given by the matrix (g_{ij}) of metric coefficients, where $g_{ij} = \langle x_i, x_j \rangle$.*

PROOF. Indeed, $I_p(x_i, x_j) = \langle x_i, x_j \rangle = g_{ij}$. □

5.7. The form I_p of plane and cylinder

Like curves, surfaces can be represented either implicitly or parametrically (see Section 3.10 for representations of curves).

The example of the sphere was discussed in Section 5.4. We now consider two more examples.

EXAMPLE 5.7.1 (Plane). The x, y-plane in $\mathbb{R}^3$ is defined implicitly by the equation $z = 0$. Consider the parametrisation $\underline{x}(u^1, u^2) = (u^1, u^2, 0) \in \mathbb{R}^3$. This is a parametrisation of the xy-plane in $\mathbb{R}^3$. Then $x_1 = (1, 0, 0)^t$, $x_2 = (0, 1, 0)^t$, and

$$g_{11} = \langle x_1, x_1 \rangle = \left\langle \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\rangle = 1,$$

and so on. Thus we have $(g_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathrm{Id}$ (the identity matrix).

EXAMPLE 5.7.2 (Cylinder). The cylinder in $\mathbb{R}^3$ is defined implicitly by the equation $x^2 + y^2 = 1$. Let $\underline{x}(u^1, u^2) = (\cos u^1, \sin u^1, u^2)$. This formula provides a parametrisation of the cylinder. We have

$$x_1 = (-\sin u^1, \cos u^1, 0)^t$$

and $x_2 = (0, 0, 1)^t$, while

$$g_{11} = \left\langle \begin{pmatrix} -\sin u^1 \\ \cos u^1 \\ 0 \end{pmatrix}, \begin{pmatrix} -\sin u^1 \\ \cos u^1 \\ 0 \end{pmatrix} \right\rangle = \sin^2 u^1 + \cos^2 u^1 = 1,$$

etc. Thus the matrix of metric coefficients is $(g_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathrm{Id}$.

REMARK 5.7.3. The two examples above illustrate that the first fundamental form does *not* contain all the information (even up to orthogonal transformations) about how the surface sits in $\mathbb{R}^3$. Indeed, the plane and the cylinder have the same first fundamental form, but are geometrically distinct embedded surfaces.

5.8. Surfaces of revolution

For surfaces of revolution, it is customary to use the notation $u^1 = \theta$ and $u^2 = \phi$. The starting point is a *generating curve* C in the xz-plane, parametrized by a pair of functions

$$x = r(\phi), \quad z = z(\phi).$$

DEFINITION 5.8.1. The surface of revolution (around the z-axis) generated by C is parametrized as follows:

$$\underline{x}(\theta, \phi) = (r(\phi)\cos\theta, r(\phi)\sin\theta, z(\phi)). \tag{5.8.1}$$

EXAMPLE 5.8.2. Consider a generating curve which is the vertical line $r(\phi) = 1$, $z(\phi) = \phi$. The resulting surface of revolution is the cylinder.

EXAMPLE 5.8.3. The generating curve $r(\phi) = \sin\phi$, $z(\phi) = \cos\phi$ yields the sphere S^2 in spherical coordinates as discussed in Section 5.4; see Example 5.8.5 for more details.

THEOREM 5.8.4. *Assume that ϕ is the arclength parameter of a parametrisation $(r(\phi), z(\phi))$ of the generating curve C. Then the first fundamental form of the corresponding surface of revolution* (5.8.1) *is given by*

$$(g_{ij}) = \begin{pmatrix} r^2(\phi) & 0 \\ 0 & 1 \end{pmatrix}.$$

PROOF. We have

$$x_1 = \frac{\partial x}{\partial \theta} = (-r\sin\theta, r\cos\theta, 0)^t,$$
$$x_2 = \frac{\partial x}{\partial \phi} = \left(\frac{dr}{d\phi}\cos\theta, \frac{dr}{d\phi}\sin\theta, \frac{dz}{d\phi}\right)^t$$

therefore

$$g_{11} = \left|\begin{pmatrix} -r\sin\theta \\ r\cos\theta \\ 0 \end{pmatrix}\right|^2 = r^2\sin^2\theta + r^2\cos^2\theta = r^2$$

and

$$
\begin{aligned}
g_{22} &= \left| \begin{pmatrix} \frac{dr}{d\phi}\cos\theta \\ \frac{dr}{d\phi}\sin\theta \\ \frac{dz}{d\phi} \end{pmatrix} \right|^2 \\
&= \left(\frac{dr}{d\phi}\right)^2 (\cos^2\theta + \sin^2\theta) + \left(\frac{dz}{d\phi}\right)^2 \\
&= \left(\frac{dr}{d\phi}\right)^2 + \left(\frac{dz}{d\phi}\right)^2
\end{aligned}
$$

and

$$
g_{12} = \left\langle \begin{pmatrix} -r\sin\theta \\ r\cos\theta \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{dr}{d\phi}\cos\theta \\ \frac{dr}{d\phi}\sin\theta \\ \frac{dz}{d\phi} \end{pmatrix} \right\rangle = -r\frac{dr}{d\phi}\sin\theta\cos\theta + r\frac{dr}{d\phi}\cos\theta\sin\theta = 0.
$$

Thus the matrix of metric coefficients is

$$
(g_{ij}) = \begin{pmatrix} r^2 & 0 \\ 0 & \left(\frac{dr}{d\phi}\right)^2 + \left(\frac{dz}{d\phi}\right)^2 \end{pmatrix}.
$$

In the case of an arclength parametrisation of the generating curve C, we obtain $g_{22} = 1$, proving the theorem. □

EXAMPLE 5.8.5. Consider the curve $(\sin\phi, \cos\phi)$ in the x, z-plane. The resulting surface of revolution is the sphere, where the ϕ parameter coincides with the angle φ of spherical coordinates. Thus, for the sphere S^2 we obtain

$$
(g_{ij}) = \begin{pmatrix} \sin^2\phi & 0 \\ 0 & 1 \end{pmatrix}.
$$

5.9. From tractrix to pseudosphere

DEFINITION 5.9.1. The *tractrix* is the curve is parametrized by $(r(\phi), z(\phi))$ where $r(\phi) = e^\phi$ and

$$
z(\phi) = \int_0^\phi \sqrt{1 - e^{2\psi}}\, d\psi = -\int_\phi^0 \sqrt{1 - e^{2\psi}}\, d\psi,
$$

where $-\infty < \phi \leq 0$.

DEFINITION 5.9.2. The *pseudosphere*[4] is the surface of revolution generated by the tractrix.

[4]The name reflects the fact that its Gaussian curvature equals -1. The Gaussian curvature of the pseudosphere will be calculated in Section 13.3.

Here $g_{11} = e^{2\phi}$, while

$$\begin{aligned} g_{22} &= (e^{\phi})^2 + \left(\sqrt{1-e^{2\phi}}\right)^2 \\ &= e^{2\phi} + 1 - e^{2\phi} = 1. \end{aligned}$$

Thus the matrix of metric coefficients is $(g_{ij}) = \begin{pmatrix} e^{2\phi} & 0 \\ 0 & 1 \end{pmatrix}$.

5.10. Chain rule in two variables

How does one measure the length of curves on a surface in terms of the metric coefficients of the surface $\underline{x}(u^1, u^2)$? Let

$$\alpha\colon [a,b] \to \mathbb{R}^2, \quad \alpha(t) = \begin{pmatrix} \alpha^1(t) \\ \alpha^2(t) \end{pmatrix}$$

be a plane curve. We will exploit the Einstein summation convention. We will also use the following version of chain rule in several variables.

THEOREM 5.10.1 (Chain rule in two variables). *Consider the curve on the surface defined by the composition $\beta(t) = \underline{x} \circ \alpha(t)$ where $\underline{x} = \underline{x}(u^1, u^2)$. Then $\frac{d\beta}{dt} = \sum_{i=1}^{2} \frac{\partial x}{\partial u^i}\frac{d\alpha^i}{dt}$, or in Einstein summation convention*

$$\frac{d\beta}{dt} = \frac{\partial \underline{x}}{\partial u^i}\frac{d\alpha^i}{dt} = J_{\underline{x}}\,\frac{d\alpha}{dt}.$$

For a proof see Keisler [**Ke74**, p. 672].

5.11. Curvature of the cusp

We provide another example of a curvature calculation. We already mentioned the cusp curve in Example 4.15.4. We will now calculate its curvature.

THEOREM 5.11.1 (Curvature of cusp curve). *Consider the cusp curve $\alpha(t) = (t^2, t^3)$ when $t \neq 0$. Then the curvature $k_\alpha(t)$ tends to infinity as we approach the cusp, i.e., as $t \to 0$.*

PROOF. The curve can be represented implicitly as the level curve of the function $F(y,x) = y^2 - x^3$: We have $F_y = 2y$, $F_x = 3x^2$, $|\nabla F| = \sqrt{4y^2 + 9x^4}$. Using the defining equation $y^2 = x^3$ we obtain

$$|\nabla F| = (4x^3 + 9x^4)^{\frac{1}{2}} = x^{\frac{3}{2}}(4+9x)^{\frac{1}{2}}. \tag{5.11.1}$$

Next, $F_{yy} = 2$, $F_{xx} = 6x$, $F_{yx} = 0$. Thus $D_B F = 2(9x^4) + 6x(4y^2) = 18x^4 + 24y^2x = 6x(3x^3 + 4y^2)$. Using the defining equation of the curve, namely $y^2 = x^3$, we obtain $D_B F = 6x(3x^3 - 4x^3) = -6x^4$. From (5.11.1) we obtain that the curvature of the curve is

$$k_\alpha = \frac{6x^4}{x^{\frac{9}{2}}(4+9x)^{3/2}} = \frac{6}{x^{1/2}(4+9x)^{3/2}}.$$

We see in particular that as we approach the cusp point (i.e., x tends to 0 or equivalently t tends to 0), the curvature tends to infinity, or in symbols $\lim_{t\to 0} k_\alpha(t) = \infty$. □

5.11.1. Curvature of the general cusp. This material is optional. We provide another example of a curvature calculation. We already mentioned the cusp curve $\alpha(t) = (t^2, t^3)$ In Example 4.15.4.

THEOREM 5.11.2. *Let* $2 < n$. *For suitable conditions on* n, *the curvature of the curve* $\alpha(t) = (t^2, t^n)$ *(where* $t \neq 0$*) tends to infinity as we approach the cusp.*

PROOF. The curve can be represented implicitly as the level curve of the function $F(y, x) = y^2 - x^n$. We have $F_y = 2y$, $F_x = nx^{n-1}$. Therefore $|\nabla F| = \sqrt{4y^2 + n^2x^{2n-2}}$. Using the defining equation of the curve, we obtain

$$|\nabla F| = (4x^n + n^2x^{2n-2})^{\frac{1}{2}} = x^{n/2}(4 + n^2x^{n-2})^{\frac{1}{2}}. \tag{5.11.2}$$

Next, $F_{yy} = 2$, $F_{xx} = n(n-1)x^{n-2}$, and $F_{yx} = 0$. Thus

$$D_B F = 2(n^2x^{2n-2}) + n(n-1)x^{n-2}(4y^2) = 2nx^{n-2}(nx^n + 2(n-1)y^2).$$

To express this as a function of a single variable, we exploit the defining equation of the curve, namely $y^2 = x^n$. Then we obtain $D_B F = 2nx^{n-2}(nx^n + (2n-2)x^n) = 2nx^{n-2}(3n-2)x^n = 2n(3n-2)x^{2n-2}$. Using (5.11.2), we obtain that the curvature of the curve is $k_\alpha = \frac{2n(3n-2)x^{2n-2}}{x^{\frac{3n}{2}}(4+n^2x^{n-2})^{\frac{3}{2}}} = x^{2n-2-\frac{3n}{2}}\left(\frac{2n(3n-2)}{(4+n^2x^{n-2})^{\frac{3}{2}}}\right) = x^{(n-4)/2}\left(\frac{2n(3n-2)}{(4+n^2x^{n-2})^{\frac{3}{2}}}\right)$. We see in particular that as we approach the cusp point (i.e., x tends to 0 or equivalently t tends to 0), the curvature tends to infinity ($\lim_{t\to 0} k_\alpha(t) = \infty$) if and only if $n < 4$. For example, this would be the case when $n = 3$ as calculated earlier. □

CHAPTER 6

Gamma symbols of a surface and applications

In this chapter we continue the study of the intrinsic geometry of surfaces in $\mathbb{R}^3$.

6.1. Measuring length of curves on surfaces

In Section 5.4, we defined the metric coefficients $g_{ij} = \langle x_i, x_j \rangle$ of a surface M with parametrisation $\underline{x}(u^1, u^2)$. The first fundamental form

$$I_p \colon T_pM \times T_pM \to \mathbb{R},$$

represented by the matrix (g_{ij}), determines the length of curves on the surface as follows.

THEOREM 6.1.1. *Consider a surface with parametrisation $\underline{x}(u^1, u^2)$. Let $\beta(t) = \underline{x} \circ \alpha(t)$, $t \in [a, b]$, be a curve on the surface. Then the length L of β is given by the formula*

$$L = \int_a^b \sqrt{g_{ij}\big(\alpha(t)\big)\frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt}}dt.$$

PROOF. The length L of β is calculated as follows using the chain rule:

$$\begin{aligned}
L &= \int_a^b \left|\frac{d\beta}{dt}\right| dt \\
&= \int_a^b \left|\sum_{i=1}^2 \frac{\partial \underline{x}}{\partial u^i}\frac{d\alpha^i}{dt}\right| dt \\
&= \int_a^b \sqrt{\left\langle \underline{x}_i \frac{d\alpha^i}{dt}, \underline{x}_j \frac{d\alpha^j}{dt}\right\rangle}dt \\
&= \int_a^b \sqrt{\langle \underline{x}_i, \underline{x}_j \rangle \frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt}}dt \\
&= \int_a^b \sqrt{g_{ij}(\alpha(t))\frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt}}dt.
\end{aligned}$$

We therefore obtain $L = \int_a^b \sqrt{g_{ij}(\alpha(t))\frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt}}dt$ as required. $\square$

Corollary 6.1.2. *If $(g_{ij}) = (\delta_{ij})$ is the identity matrix, then the length of the curve $\beta = \underline{x} \circ \alpha$ on the surface equals the length of the original curve α in $\mathbb{R}^2$.*

Proof. If the first fundamental form of the surface satisfies $g_{ij} = \delta_{ij}$, then $L = \int_a^b \sqrt{\left(\frac{d\alpha^1}{dt}\right)^2 + \left(\frac{d\alpha^2}{dt}\right)^2}dt$. By Lemma 4.15.2, this is the length of the curve $\alpha(t)$ in $\mathbb{R}^2$.[1] □

6.2. Normal vector to a surface

We first review the notation introduced in Section 5.2.

- $\underline{x}\colon \mathbb{R}^2 \to \mathbb{R}^3$ is a regular parametrisation of a surface $M \subseteq \mathbb{R}^3$;
- $x_i = \frac{\partial \underline{x}}{\partial u^i}$, $i = 1, 2$ are its tangent vectors;
- the tangent vectors span the tangent plane $T_p = \operatorname{Span}(x_1, x_2)$ at a point $p = \underline{x}(u^1, u^2)$ of the surface.

Definition 6.2.1. The normal vector $n(u^1, u^2)$ to a regular surface at the point $\underline{x}(u^1, u^2) \in \mathbb{R}^3$ is defined in terms of the vector product, *cf.* Definition 1.10.1, as follows:

$$n = \frac{x_1 \times x_2}{|x_1 \times x_2|},$$

so that $\langle n, x_i \rangle = 0$ for each $i = 1, 2$.

Remark 6.2.2 (Sign ambiguity). Either the vector $\frac{x_1 \times x_2}{|x_1 \times x_2|}$ or the opposite vector

$$\frac{x_2 \times x_1}{|x_1 \times x_2|} = -\frac{x_1 \times x_2}{|x_1 \times x_2|}$$

can be taken to be a normal vector to the surface.

Theorem 6.2.3. *Consider the surface $M \subseteq \mathbb{R}^3$ given by the graph $z = f(x, y)$ of a function $f(x, y)$. Then the vector*

$$(f_x, f_y, -1)^t$$

is orthogonal to the tangent plane T_p to M at the point $p = (x, y, f(x, y))$.

Proof. A standard parametrisation of the graph of f is

$$\underline{x}(u^1, u^2) = (u^1, u^2, f(u^1, u^2)).$$

[1] To simplify notation, let $x = x(t) = \alpha^1(t)$ and $y = y(t) = \alpha^2(t)$. Then the dt cancels out. The infinitesimal element of arclength is $ds = \sqrt{dx^2 + dy^2}$. The length of the curve is $\int ds = \int \sqrt{dx^2 + dy^2}$.

Then $x_1 = (1, 0, f_x)^t$ and $x_2 = (0, 1, f_y)^t$. Taking the cross product, we obtain

$$\det \begin{pmatrix} \overrightarrow{i} & \overrightarrow{j} & \overrightarrow{k} \\ 1 & 0 & f_x \\ 0 & 1 & f_y \end{pmatrix} = -f_x \overrightarrow{i} - f_y \overrightarrow{j} + \overrightarrow{k}$$

which is the opposite of the vector $(f_x, f_y, -1)^t$. Note that this may not be a unit vector. □

COROLLARY 6.2.4. *The normal vector n of the graph of $f(x, y)$ in $\mathbb{R}^3$ is*

$$n = \frac{(f_x, f_y, -1)^t}{\sqrt{f_x^2 + f_y^2 + 1}}$$

up to sign.

PROOF. This is immediate from the theorem by normalizing the vector $(f_x, f_y, -1)^t$. □

6.3. Γ symbols of a surface

The symbols Γ_{ij}^k, roughly speaking,[2] account for how the surface twists[3] in space. They are, however, coordinate dependent (unlike the Gaussian curvature that we will define in Section 9.5).

REMARK 6.3.1 (Relation to the geodesic equation). We will see that the symbols Γ_{ij}^k also control the behavior of geodesics on the surface. Here geodesics on a surface can be thought of as curves that are to the surface what straight lines are to a plane, or what great circles are to a sphere; *cf.* Definition 7.2.2.

REMARK 6.3.2. The metric coefficients g_{ij} were defined in terms of *first* partial derivatives of $\underline{x}$. Meanwhile, the symbols Γ_{ij}^k are defined in terms of the *second* partial derivatives.[4]

The following is immediate from the definition of regularity.

LEMMA 6.3.3. *For a regular parametrisation $\underline{x}$, the vectors (x_1, x_2, n) form a basis (frame) for $\mathbb{R}^3$.*

The symbols Γ are defined as the coefficients of the decomposition of the second partial derivative vector x_{ij}, defined by

$$x_{ij} = \frac{\partial^2 \underline{x}}{\partial u^i \partial u^j}$$

[2]In the first approximation: kiruv rishon.

[3]Mitpatelet

[4]Since the Gamma symbols do not transform as a tensor, one does not stagger the indices.

with respect to the basis, or frame,[5] (x_1, x_2, n), as follows.

DEFINITION 6.3.4. The symbols Γ_{ij}^k are uniquely determined by the decomposition

$$x_{ij} = \Gamma_{ij}^1\, x_1 + \Gamma_{ij}^2\, x_2 + L_{ij} n$$

REMARK 6.3.5. The coefficients L_{ij} are also uniquely defined by the formula. Their geometric meaning will be analyzed in Section 9.6.

6.4. Basic properties of the Γ symbols

PROPOSITION 6.4.1. *We have the following formula for the Gamma symbols:*

$$\Gamma_{ij}^k = \langle x_{ij}, x_\ell \rangle g^{\ell k}, \tag{6.4.1}$$

where $(g^{\ell k})$ is the inverse matrix of (g_{ij}).

PROOF. We have

$$\begin{aligned} \langle x_{ij}, x_\ell \rangle &= \langle \Gamma_{ij}^k x_k + L_{ij} n, x_\ell \rangle \\ &= \langle \Gamma_{ij}^k x_k, x_\ell \rangle + \langle L_{ij} n, x_\ell \rangle \\ &= \Gamma_{ij}^k g_{k\ell} \end{aligned}$$

since the vector n is perpendicular to each tangent vector of the surface. We now multiply by $g^{\ell m}$ and sum over the index ℓ:

$$\langle x_{ij}, x_\ell \rangle g^{\ell m} = \Gamma_{ij}^k g_{k\ell} g^{\ell m} = \Gamma_{ij}^k \delta_k^m = \Gamma_{ij}^m.$$

This is equivalent to the desired formula. □

REMARK 6.4.2. We have the following relation: $\Gamma_{ij}^k = \Gamma_{ji}^k$, or $\Gamma_{[ij]}^k = 0$ for all i, j, k.

THEOREM 6.4.3. *For the standard parametrisations of the plane and the cylinder, the Gamma symbols vanish.*

PROOF. We calculate the symbols Γ_{ij}^k for the plane $\underline{x}(u^1, u^2) = (u^1, u^2, 0)$. We have $x_1 = (1, 0, 0)^t$, $x_2 = (0, 1, 0)^t$. Thus we have $x_{ij} = 0$ for all i, j. Hence by formula (6.4.1), $\Gamma_{ij}^k = 0$ for all i, j, k.

Next, we calculate the symbols for the cylinder with parametrisation $\underline{x}(u^1, u^2) = (\cos u^1, \sin u^1, u^2)$. We obtain the following data:

$$\begin{cases} x_1 = (-\sin u^1, \cos u^1, 0)^t; \\ x_2 = (0, 0, 1)^t; \\ n = (\cos u^1, \sin u^1, 0)^t. \end{cases}$$

[5] Maarechet yichus

Thus we have $x_{22} = 0$, $x_{21} = 0$ and so $\Gamma^k_{22} = 0$ and $\Gamma^k_{12} = \Gamma^k_{21} = 0$ for $k = 1, 2$.

Meanwhile, $x_{11} = (-\cos u^1, -\sin u^1, 0)^t$. This vector is proportional to n:

$$x_{11} = 0x_1 + 0x_2 + (-1)n.$$

Hence $\Gamma^k_{11} = 0$ for all k. □

An example of a surface with nonzero Γ symbols will appear in Section 6.7.

6.5. Derivatives of the metric coefficients

DEFINITION 6.5.1. We will use the following *comma notation* for the partial derivative of the function $g_{ij} = g_{ij}(u^1, u^2)$:

$$g_{ij,k} = \frac{\partial}{\partial u^k}(g_{ij}).$$

LEMMA 6.5.2 (Leibniz rule). *Scalar product of vector valued functions f, g satisfies Leibniz's rule:*

$$\langle f, g\rangle' = \langle f', g\rangle + \langle f, g'\rangle.^6 \tag{6.5.1}$$

PROPOSITION 6.5.3. *In terms of the symmetrisation notation introduced in Section* 1.6*, we have*

$$g_{ij,k} = 2g_{m\{i}\Gamma^m_{j\}k}, \tag{6.5.2}$$

or more explicitly

$$g_{ij,k} = g_{mi}\Gamma^m_{jk} + g_{mj}\Gamma^m_{ik}. \tag{6.5.3}$$

PROOF. Indeed, by Leibniz rule (Lemma 6.5.2),

$$\begin{aligned} g_{ij,k} &= \frac{\partial}{\partial u^k}\langle x_i, x_j\rangle \\ &= \langle x_{ik}, x_j\rangle + \langle x_i, x_{jk}\rangle \\ &= \langle \Gamma^m_{ik}x_m, x_j\rangle + \langle \Gamma^m_{jk}x_m, x_i\rangle \\ &= \Gamma^m_{ik}\langle x_m, x_j\rangle + \Gamma^m_{jk}\langle x_m, x_i\rangle. \end{aligned}$$

[6] For a proof, see infi 3 or [**Leib**]. In more detail, let (f_1, f_2) be components of f, and let (g_1, g_2) be components of g. Then $\langle f, g\rangle' = (f_1g_1 + f_2g_2)' = f_1g_1' + f_1'g_1 + f_2g_2' + f_2'g_2 = \langle f', g\rangle + \langle f, g'\rangle$. The same proof goes through for arbitrary number of components, e.g., for functions with values in $\mathbb{R}^3$.

The scalar products above by definition are the metric coefficients. Therefore we have

$$\begin{aligned} g_{ij,k} &= \Gamma_{ik}^m g_{mj} + \Gamma_{jk}^m g_{mi} \\ &= g_{mj}\Gamma_{ik}^m + g_{mi}\Gamma_{jk}^m \\ &= 2g_{m\{j}\Gamma_{i\}k}^m \\ &= 2g_{m\{i}\Gamma_{j\}k}^m \end{aligned}$$

as required. □

Remark 6.5.4. The system of equations of type (6.5.3) suggests that one may be able to solve the system for Γ_{ij}^k, i.e., to express Γ_{ij}^k in terms of the metric coefficients g_{ij} and their derivatives $g_{ij,k}$. This turns out to be correct, as we show in Section 6.6.

6.6. Intrinsic nature of the Γ symbols

The coefficients Γ are intrinsic[7] meaning that they are determined by the metric coefficients alone, and are therefore independent of the ambient (extrinsic) geometry of the surface, i.e., the way the surface "sits" in 3-space.

Theorem 6.6.1. *The symbols Γ_{ij}^k can be expressed in terms of the first fundamental form and its derivatives as follows:*

$$\Gamma_{ij}^k = \frac{1}{2}\left(g_{i\ell,j} - g_{ij,\ell} + g_{j\ell,i}\right) g^{\ell k},$$

where g^{ij} is the inverse matrix of g_{ij}.

Proof. The proof is a calculation. Applying Proposition 6.5.3 three times, we obtain

$$\begin{aligned} g_{i\ell,j} - g_{ij,\ell} + g_{j\ell,i} &= 2g_{m\{i}\Gamma_{\ell\}j}^m - 2g_{m\{i}\Gamma_{j\}\ell}^m + 2g_{m\{j}\Gamma_{\ell\}i}^m \\ &= g_{mi}\Gamma_{\ell j}^m + g_{m\ell}\Gamma_{ij}^m - g_{mi}\Gamma_{j\ell}^m - g_{mj}\Gamma_{i\ell}^m + g_{mj}\Gamma_{\ell i}^m + g_{m\ell}\Gamma_{ji}^m \\ &= 2g_{m\ell}\Gamma_{ji}^m. \end{aligned}$$

Thus $\frac{1}{2}(g_{i\ell,j} - g_{ij,\ell} + g_{j\ell,i})g^{\ell k} = \Gamma_{ij}^m g_{m\ell} g^{\ell k} = \Gamma_{ij}^m \delta_m^k = \Gamma_{ij}^k$, as required.[8] □

---

[7]This is atzmit and not pnimit according to Vishne.

[8]Another approach would be to prove the identity $2\langle x_{ij}, x_\ell\rangle = g_{i\ell,j} - g_{ij,\ell} + g_{j\ell,i}$ and use it to prove the theorem.

Γ_{ij}^1	$j=1$	$j=2$
$i=1$	0	$\frac{1}{r}\frac{dr}{d\phi}$
$i=2$	$\frac{1}{r}\frac{dr}{d\phi}$	0

TABLE 6.7.1. Symbols Γ_{ij}^1 of a surface of revolution (6.7.1)

6.7. Gamma symbols for a surface of revolution

In Section 5.8, we defined a surface of revolution as obtained by starting with a curve $(r(\phi), z(\phi))$ in the (x, z) plane, and rotating it around the z-axis. This produces the parametrisation $\underline{x}(\theta, \phi) = (r(\phi)\cos\theta, r(\phi)\sin\theta, z(\phi))$. Here we adopt the notation

$$u^1 = \theta, \quad u^2 = \phi.$$

THEOREM 6.7.1. *For a surface of revolution we have* $\Gamma_{11}^1 = \Gamma_{22}^1 = 0$, *while* $\Gamma_{12}^1 = \frac{1}{r}\frac{dr}{d\phi}$, *cf. Table* 6.7.1.

PROOF. For the surface of revolution

$$\underline{x}(\theta, \phi) = (r(\phi)\cos\theta, r(\phi)\sin\theta, z(\phi)), \tag{6.7.1}$$

the metric coefficients are given by the matrix

$$(g_{ij}) = \begin{pmatrix} r^2(\phi) & 0 \\ 0 & \left(\frac{dr}{d\phi}\right)^2 + \left(\frac{dz}{d\phi}\right)^2 \end{pmatrix}$$

by Theorem 5.8.4. We will use the formula $\Gamma_{ij}^k = \frac{1}{2}(g_{i\ell,j} - g_{ij,\ell} + g_{j\ell,i})g^{\ell k}$ from Theorem 6.6.1. Since the off-diagonal coefficient $g_{12} = 0$ vanishes, the diagonal coefficients of the inverse matrix satisfy

$$g^{ii} = \frac{1}{g_{ii}} \text{ for each } i. \tag{6.7.2}$$

We have $\dfrac{\partial}{\partial\theta}(g_{ii}) = 0$ since the coefficients g_{ii} depend only on the variable ϕ. Thus the terms

$$g_{ii,1} = 0 \tag{6.7.3}$$

vanish. Let us now compute the symbols Γ_{ij}^1 for $k = 1$. Using formulas (6.7.2) and (6.7.3), we obtain

$$\begin{aligned} \Gamma_{11}^1 &= \frac{1}{2g_{11}}(g_{11,1} - g_{11,1} + g_{11,1}) && \text{by formula (6.7.2)} \\ &= 0 && \text{by formula (6.7.3).} \end{aligned}$$

Similarly, $\Gamma^1_{22} = \frac{1}{2g_{11}}(g_{12,2} - g_{22,1} + g_{12,2}) = \frac{g_{12,2}}{g_{11}} = \frac{d}{d\phi}(0) = 0$. Finally,

$$\begin{aligned}\Gamma^1_{12} &= \frac{1}{2g_{11}}(g_{11,2} - g_{12,1} + g_{12,1}) = \frac{g_{11,2}}{2g_{11}} \\ &= \frac{\frac{dr}{d\phi}}{r},\end{aligned}$$

proving the theorem.[9] $\square$

6.8. Spherical coordinates

Spherical coordinates are useful in understanding surfaces of revolution. They were already introduced in Section 5.4.

DEFINITION 6.8.1. Spherical coordinates (ρ, θ, φ) in $\mathbb{R}^3$ are defined by the following formulas. We have

$$\rho = \sqrt{x^2 + y^2 + z^2} = \sqrt{r^2 + z^2}$$

is the distance to the origin, $r = \sqrt{x^2 + y^2}$, while φ is the angle with the z-axis, so that $\cos\varphi = \frac{z}{\rho}$. Here θ is the angle inherited from polar coordinates in the x, y plane, so that $\tan\theta = \frac{y}{x}$ while $x = r\cos\theta$ and $y = r\sin\theta$.

REMARK 6.8.2. The interval of definition for the variable φ is $\varphi \in [0, \pi]$ since $\varphi = \arccos\frac{z}{\rho}$ and the range of the arccos function is $[0, \pi]$. Meanwhile $\theta \in [0, 2\pi]$ as usual.

EXAMPLE 6.8.3. The unit sphere $S^2 \subseteq \mathbb{R}^3$ is defined in spherical coordinates by the condition

$$S^2 = \{(\rho, \theta, \varphi) \colon \rho = 1\}.$$

DEFINITION 6.8.4. A *latitude*[10] on the unit sphere is a circle satisfying the equation

$$\varphi = \text{constant}.$$

DEFINITION 6.8.5. The *equator* of the sphere is defined by the equation $\{\varphi = \frac{\pi}{2}\}$.

The equator is the only latitude that is also a great circle (see Section 6.9) of the sphere.

Each latitude of the sphere is parallel to the equator (i.e., lies in a plane parallel to the plane of the equator). A latitude can be parametrized by setting $\theta(t) = t$ and $\varphi(t) = \text{constant}$.

---

[9] To keep track of the variables θ and ϕ, one could use the notation $\Gamma^\theta_{\theta\phi}$ for Γ^1_{12}, etc. Then one would write $\Gamma^\theta_{\theta\phi} = \frac{\frac{dr}{d\phi}}{r}$, whereas $\Gamma^\theta_{\theta\theta} = \Gamma^\theta_{\phi\phi} = 0$.

[10] kav rochav

LEMMA 6.8.6. *On the unit sphere defined by equation $\{\rho = 1\}$, at each point we have the relation*

$$r = \sin\varphi, \tag{6.8.1}$$

where r is the distance from the point to the z-axis.

This is immediate from the relation $r = \rho \sin\varphi$.

6.9. Great circles

In Section 8.2, we will analyze the geodesic equation. This is a system of nonlinear second order differential equations. Our goal in this section is to provide a geometric intuition for this equation. We will establish a connection between the following two items:

(1) solutions of this system of ODEs, and
(2) great circles on the sphere.

REMARK 6.9.1 (Clairaut's relation as a bridge). Such a connection is established via the intermediary of Clairaut's relation for a variable point q on a great circle:

$$r(t)\cos\gamma(t) = \text{const}$$

where r is the distance from q to the (vertical) axis of revolution and γ is the angle at q between the direction of the great circle and the latitudinal circle; *cf.* Theorem 6.10.4.

REMARK 6.9.2. In Section 6.10.4, we will verify Clairaut's relation for great circles using synthetic geometry, and also show that the great circles satisfy a first order differential equation. In Section 8.2 we will complete the connection by deriving Clairaut's relation from the geodesic equation. Thus Clairaut's relation provides a bridge between geometry and analysis.

Let $S^2 \subseteq \mathbb{R}^3$ be the unit sphere defined by the equation

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

DEFINITION 6.9.3. A plane P through the origin is given by an equation

$$ax + by + cz = 0, \tag{6.9.1}$$

where $a^2 + b^2 + c^2 > 0$.

DEFINITION 6.9.4. A great circle G of S^2 is given by an intersection

$$G = S^2 \cap P.$$

Example 6.9.5 (A parametrisation of the equator of S^2). The equator is parametrized by

$$\alpha(t) = (\cos t)e_1 + (\sin t)e_2.$$

Theorem 6.9.6. *Every great circle can be parametrized by*

$$\alpha(t) = (\cos t)v + (\sin t)w$$

where $v, w \in S^2$ are orthonormal vectors in $\mathbb{R}^3$.

Proof. By the Pythagorean theorem, $\alpha(t)$ is a unit vector and therefore lies on the unit sphere. $\square$

Example 6.9.7 (an implicit (non-parametric) representation of a great circle). Recall that we have

$$x = r\cos\theta = \rho\sin\varphi\cos\theta;\ y = \rho\sin\varphi\sin\theta;\ z = \rho\cos\varphi\,. \tag{6.9.2}$$

If the circle lies in the plane $ax+by+cz=0$ where a, b, c are fixed, the great circle in coordinates (θ, φ) is defined implicitly by the equation

$$a\sin\varphi\cos\theta + b\sin\varphi\sin\theta + c\cos\varphi = 0,$$

as in (6.9.1).

6.10. Position vector orthogonal to tangent vector of S^2

Lemma 6.10.1. *Let $\alpha\colon \mathbb{R} \to S^2$ be a parametrized curve on the sphere $S^2 \subseteq \mathbb{R}^3$. Then the tangent vector $\dfrac{d\alpha}{dt}$ is perpendicular to the position vector $\alpha(t)$, or in formulas:*

$$\left\langle \alpha(t), \frac{d\alpha}{dt} \right\rangle = 0.$$

Proof. We have $\langle \alpha(t), \alpha(t)\rangle = 1$ by definition of S^2. We apply the operator $\frac{d}{dt}$ to obtain

$$\frac{d}{dt}\langle \alpha(t), \alpha(t)\rangle = 0. \tag{6.10.1}$$

Next, we apply the Leibniz rule (6.5.1) to equation (6.10.1) to obtain

$$\left\langle \alpha(t), \frac{d\alpha}{dt} \right\rangle + \left\langle \frac{d\alpha}{dt}, \alpha(t) \right\rangle = 2\left\langle \alpha(t), \frac{d\alpha}{dt} \right\rangle = 0,$$

completing the proof. $\square$

The lemma will be used in the analysis of Clairaut's relation in Section 6.10.1.

6.10.1. Clairaut's relation: bridge from geometry to analysis. This material is optional. Our interest in Clairaut's relation (6.10.2) for curves on the sphere (and more general surfaces of revolution) lies in the motivation it provides for the general geodesic equation on a surface of revolution, as a bridge between geometry and analysis.

REMARK 6.10.2. We will use Newton's dot notation for the derivative with respect to t as in $\dot{\alpha}(t)$, where t is not necessarily an arclength parameter.

We defined great circles in Section 6.9. Consider a great circle $G \subseteq S^2$. Let $\alpha(t)$ be a parametrisation of G.

DEFINITION 6.10.3 (Angle γ with latitude). $\gamma(t)$ is the angle, at the point $\alpha(t) \in S^2$, between the tangent vector $\dot{\alpha}(t)$ to the curve and the vector tangent to the latitude passing through $\alpha(t)$.

THEOREM 6.10.4 (Clairaut's relation). *Let $\alpha(t)$ be a regular parametrisation of a great circle G on $S^2 \subseteq \mathbb{R}^3$. Let $r(t)$ denote the distance from the point $\alpha(t)$ to the z-axis. Then the following relation holds along G:*

$$r(t)\cos\gamma(t) = \text{const.} \tag{6.10.2}$$

Here the constant has value const $= r_{\min}$, *where $r_{\min}$ is the least Euclidean distance from a point of G to the z-axis.*

The proof exploits the sine law of spherical trigonometry given in Section 6.10.2.

6.10.2. Spherical sine law. Let S^2 be the unit sphere in 3-space.

DEFINITION 6.10.5. A *spherical triangle* is the following collection of data:

(1) the three vertices are points of S^2, not lying on a common great circle;
(2) each side is an arc of of a great circle;
(3) the angle at each vertex is the angle between tangent vectors to the sides;
(4) each side has length strictly smaller than π.

THEOREM 6.10.6 (Spherical sine law). *Let a be the side opposide angle α, let b be the side opposite angle β, and let c be the side opposite angle γ. Then*

$$\frac{\sin a}{\sin\alpha} = \frac{\sin b}{\sin\beta} = \frac{\sin c}{\sin\gamma}\,.$$

COROLLARY 6.10.7. *If $\gamma = \frac{\pi}{2}$ then*

$$\sin a = \sin c\,\sin\alpha. \tag{6.10.3}$$

REMARK 6.10.8. For small values of the sides a, c in a right-angle triangle, we recapture the Euclidean sine law

$$a = c\,\sin\alpha$$

as the limiting case of (6.10.3).

Three proofs of the spherical sine law are given in Subsection 6.10.7.

6.10.3. Longitudes; length-minimizing property.

DEFINITION 6.10.9. A longitude[11] on S^2 is the arc of a great circle connecting the North Pole $e_3 = (0,0,1)^t$ and the South Pole $-e_3 = (-1,0,0)^t$.

Next we show that the arc of great circle is length-minimizing among all paths joining a pair of points on the sphere. Recall that the metric coefficients of the sphere in coordinates (θ, ϕ) are

$$\begin{cases} g_{11} = \sin^2\phi, \\ g_{22} = 1, \\ g_{12} = 0. \end{cases}$$

THEOREM 6.10.10. *A shortest spherical path between a pair of points on S^2 is an arc of great circle passing through them.*

PROOF. Let $p_0, p_1 \in S^2$.

Step 1. Since orthogonal transformations preserve lengths, by applying a rotation we can assume that p_0 and p_1 lie on a common longitude. To fix ideas, we will assume that this is the longitude in the xz-plane defined by the coordinate $\theta = 0$ while ϕ varies from 0 to π.

Step 2. Let $\underline{x}(\theta, \phi)$ be the standard parametrisation of the sphere as a surface of revolution. Then $p_i = \underline{x}(0, \phi_i)$, for $i = 0, 1$.

Step 3. The main point is that moving in the latitudinal direction can only increase the length of the curve. To implement this technically, consider an arbitrary path $\alpha(t)$, $t \in [0,1]$ joining the two points, so that $\alpha(0) = (0, \phi_0)$ and $\alpha(1) = (0, \phi_1)$. Let $\beta(t) = \underline{x}(\alpha(t))$ be the path on the sphere. Then $p_i = \beta(i)$ for $i = 0, 1$. The length L of the path β can now be bounded from below as follows using Theorem 6.1.1:

$$\begin{aligned} L = \int_0^1 \left|\frac{d\beta}{dt}\right| &= \int_0^1 \sqrt{\sin^2\phi \left(\tfrac{d\theta}{dt}\right)^2 + \left(\tfrac{d\phi}{dt}\right)^2} \\ &\geq \int_0^1 \left|\frac{d\phi}{dt}\right| \\ &\geq \int_0^1 \frac{d\phi}{dt} \\ &= \int_{\phi_0}^{\phi_1} d\phi = \phi_1 - \phi_0. \end{aligned}$$

Step 4. The difference $\phi_1 - \phi_0$ is precisely the length of the segment of the longitude between the two points. This proves that the arc of the longitude containing both points is a shortest path between them. □

[11]kav orech

A proof without calculus is given at `https://matheducators.stackexchange.com/a/26698/1385`.

DEFINITION 6.10.11. A great circle will be called *generic* if it does not pass through the north and south poles, i.e., does not include a longitude.

DEFINITION 6.10.12. Let $p_G \in S^2$ be the point of the great circle G with the largest z-coordinate among all points of G.

LEMMA 6.10.13. *Consider a generic great circle $G \subseteq S^2$. Let $\alpha(t)$ be a regular parametrisation of G with $\alpha(0) = p_G$, and denote by $\dot\alpha(0)$ the tangent vector to G at p_G. Then G has the following three equivalent properties:*

(1) $\dot\alpha(0)$ *is proportional to the vector product $e_3 \times p_G$;*
(2) $\dot\alpha(0)$ *is perpendicular to the longitude passing through p_G;*
(3) $\dot\alpha(0)$ *is tangent at p_G to the latitude.*

PROOF. We will prove item (1).

Step 1. Note that $\langle \dot\alpha(0), p_G\rangle = 0$ by Lemma 6.10.1 (tangent vector is perpendicular to the position vector of the point on the sphere). It remains to prove that $\dot\alpha(0)$ is orthogonal to e_3.

Step 2. Since p_G as the point with maximal z-coordinate, the function $\langle \alpha(t), e_3\rangle$ achieves its maximum at $t = 0$. Hence by Fermat's theorem we have

$$\frac{d}{dt}\Big|_{t=0} \langle \alpha(t), e_3\rangle = 0. \tag{6.10.4}$$

Step 3. We apply the Leibniz rule to (6.10.4) to obtain

$$\left\langle \frac{d\alpha}{dt}\Big|_{t=0}, e_3 \right\rangle = -\left\langle \alpha(t), \frac{de_3}{dt} \right\rangle = 0,$$

since e_3 is constant. Thus $\langle \dot\alpha(0), e_3\rangle = 0$.[12] □

6.10.4. Preliminaries to proof of Clairaut's relation.

DEFINITION 6.10.14. The spherical distance $d(p, q)$ on S^2 is the distance between points $p, q \in S^2$ measured along shortest arc of great circle passing through them:

$$d(p, q) = \arccos\langle p, q\rangle.$$

[12]**Equation of great circle.** This material is optional. One can show that a great circle on the sphere satisfies the equation $\cot(\varphi) = \tan(\gamma)\cos(\theta - \theta_0)$, where γ is the angle of inclination of the plane (see below). Indeed, a great circle is the intersection of the sphere with a plane through the origin. Let a unit normal to that plane be $\mathbf{u} = [-\sin(\gamma), 0, \cos(\gamma)]$, where for convenience we choose our x and y axes so that $y = 0$ (the plane contains the y-axis). Then the equation $\mathbf{u} \cdot [\sin\varphi\cos\theta, \sin\varphi\sin\theta, \cos\varphi] = 0$ becomes $\cot(\varphi) = \tan(\gamma)\cos(\theta)$. Rotating around the z axis, this becomes $\cot(\varphi) = \tan(\gamma)\cos(\theta - \theta_0)$ (this gamma has nothing to do with the gamma from Clairaut's relation).

By Theorem 6.10.10, $d(p,q)$ is the least length of a path on the sphere joining p,q.

LEMMA 6.10.15. *When p be the north pole $p=e_3$, the spherical distance $d(e_3,q)$ is the φ-coordinate of the point q.*

PROOF. This follows from the fact that the length of an arc of a unit circle equals the subtended angle. □

As in Section 6.10.1, we denote by $\gamma(t)$ the angle formed by the tangent vector to a great circle and the latitude. First we note that the complementary angle of γ is the angle with the longitude, as follows.

LEMMA 6.10.16. *Let $\alpha(t)$ be a regular parametrisation of a great circle. Then the following are equivalent:*

(1) *the angle between $\dot\alpha(t)$ and (the vector tangent to) the latitude is $\gamma(t)$;*
(2) *the angle between $\dot\alpha(t)$ and the vector tangent to the longitude at the point $\alpha(t)$ is of $\frac{\pi}{2}-\gamma(t)$.*

PROOF. For a surface of revolution $\underline{x}(\theta,\varphi)$ (where $u^1=\theta$ and $u^2=\varphi$), the metric coefficient g_{12} vanishes (see Section 5.6 and Section 6.7). Hence the tangent vectors $x_1=\frac{\partial x}{\partial\theta}$ and $x_2=\frac{\partial x}{\partial\varphi}$ are orthogonal. These are respectively the tangents to the latitude and the longitude. Therefore the two angles add up to $\frac{\pi}{2}$. □

6.10.5. Proof of Clairaut's relation.

Let us summarize our notation for a great circle $G\subseteq S^2$ with parametrisation $\beta(t)$.

(1) $p_G\in S^2$ is the point of G with maximal z-coordinate.
(2) $\varphi(t)$ is the spherical coordinate φ at the point $\beta(t)$.
(3) $r(t)=\sin\varphi(t)$ is the distance to the z-axis.
(4) Δ_t is a spherical triangle with vertices at $\beta(t)$, p_G, and the north pole e_3.
(5) The angle of the triangle Δ_t at the vertex p_G is $\frac{\pi}{2}$ by Lemma 6.10.13.

We introduce two further arcs:

- c_t is the arc of longitude joining the variable point $\beta(t)$ to e_3.
- b is the arc of longitude joining p_G to e_3.

By Lemma 6.10.15, the lengths b and c_t are respectively the φ-coordinates of the points p_G and $\beta(t)$. We now apply Corollary 6.10.7 (of the law of sines) to the right-angle triangle Δ_t. We obtain

$$\sin c_t \sin\left(\tfrac{\pi}{2}-\gamma(t)\right)=\sin b.$$

Note that $r(t)=\sin c_t$ by (6.8.1). Since b is independent of t, we obtain the relation $r(t)\cos\gamma(t)=\sin b$, proving Clairaut's relation.

6.10.6. Differential equation of great circle. Recall that the sphere $S^2 \subseteq \mathbb{R}^3$ is defined in spherical coordinates (θ, φ, ρ) by the equation $\rho = 1$. Note that the parametrisation of the great circle G in Clairaut's relation need not be arclength. Assume that a great circle $G \subseteq S^2$ is generic (see Definition 6.10.11). Then we can parametrize G by the value of the spherical coordinate θ. By the implicit function theorem, we obtain the following lemma.

LEMMA 6.10.17. *A generic great circle $G \subseteq S^2$ can be defined by a suitable function $\varphi = \varphi(\theta)$.*

THEOREM 6.10.18. *Each generic great circle $G \subseteq S^2$ satisfies the following differential equation for $\varphi = \varphi(\theta)$:*

$$r^2 + \left(\frac{d\,\varphi}{d\theta}\right)^2 = \frac{r^4}{\mathrm{const}^2}, \tag{6.10.5}$$

where $r = \sin\varphi$ and $\mathrm{const} = \sin\varphi_{\min}$ *from Theorem* 6.10.4.

PROOF. An *element of length*, denoted ds, along the great circle G decomposes into a longitudinal (along a longitude, north-south) displacement $d\varphi$, and a latitudinal (east-west) displacement $rd\theta$ (Leibniz's characteristic triangle). Thus $ds^2 = r^2 d\theta^2 + d\varphi^2$ (for details see Proposition 8.4.1). Recall that γ is the angle with the latitude. Hence

$$\begin{cases} ds \sin\gamma = d\varphi \\ ds\cos\gamma = rd\theta \end{cases}$$

along G. Dividing the first equation by the second, we obtain

$$\tan\gamma = \frac{\sin\gamma}{\cos\gamma} = \frac{d\varphi}{r\,d\theta}.$$

Expressing $\cos^2\gamma$ in terms of $\tan^2\gamma$, we obtain

$$\cos^2\gamma = \frac{1}{1 + \left(\frac{d\varphi}{r\,d\theta}\right)^2}. \tag{6.10.6}$$

Substituting into (6.10.6) the value $\cos\gamma = \frac{\mathrm{const}}{r}$ from Clairaut's relation (Theorem 6.10.4), we obtain[13]

$$\left(\frac{\mathrm{const}}{r}\right)^2 \left(1 + \left(\frac{d\varphi}{rd\theta}\right)^2\right) = 1. \tag{6.10.8}$$

[13] Equivalently, $1 + \left(\frac{d\varphi}{rd\theta}\right)^2 = \left(\frac{r}{\mathrm{const}}\right)^2$ or $\left(\frac{d\varphi}{rd\theta}\right)^2 = \frac{r^2}{\mathrm{const}^2} - 1$ or $\frac{d\varphi}{rd\theta} = \sqrt{\frac{r^2}{\mathrm{const}^2} - 1}$ or $\frac{1}{r}\frac{d\varphi}{d\theta} = \sqrt{\frac{r^2}{\mathrm{const}^2} - 1}$ or

$$\frac{1}{\sin\varphi}\frac{d\varphi}{d\theta} = \sqrt{\frac{\sin^2\varphi}{\mathrm{const}^2} - 1} \tag{6.10.7}$$

Multiplying by r^4 and dividing by const^2 we obtain

$$r^2 + \left(\frac{d\varphi}{d\theta}\right)^2 = \frac{r^4}{\text{const}^2} \qquad \text{where} \quad r = \sin\varphi,$$

proving the theorem. □

This equation is solved explicitly in terms of integrals in note 4 below.[14]

6.10.7. Sine law. This material is optional. We present three proofs.

First proof of spherical sine law. Let A, B, C be the angles. We choose a coordinate system so that the three vertices of the spherical triangle are located at $(1,0,0)$, $(\cos a, \sin a, 0)$ and $(\cos b, \sin b \cos C, \sin b \sin C)$. The volume of the tetrahedron formed from these three vertices and the origin is $\frac{1}{6}\sin a \sin b \sin C$. Since this volume is invariant under cyclic relabeling of the sides and angles, we have $\sin a \sin b \sin C = \sin b \sin c \sin A = \sin c \sin a \sin B$ and therefore $\frac{\sin A}{\sin a} = \frac{\sin B}{\sin b} = \frac{\sin C}{\sin c}$ proving the law. See `http://math.stackexchange.com/questions/1735860`. □

Second proof. Denote by A, B, C the vertices opposite the sides a, b, c. The points A, B, C can be thought of as unit vectors in $\mathbb{R}^3$. Note that $\sin a = |B \times C|$, etc. Meanwhile $\sin\alpha = |C' \times B'|$ where C' is normalisation of the orthogonal component of C when the component parallel to A is eliminated (through the Gram-Schmidt process). Here the component of C parallel to A is $(C \cdot A)A$ of norm $\cos b$. The orthogonal component is $C - (C \cdot A)A$ is of norm $\sin b = |A \times C| = |C - (C \cdot A)A|$. Thus we have $C' = \frac{C-(C\cdot A)A}{|A\times C|}$, $\quad B' = \frac{B-(B\cdot A)A}{|A\times B|}$. Therefore $\frac{\sin\alpha}{\sin a} = \frac{\left|\frac{C-(C\cdot A)A}{|A\times C|}\times\frac{B-(B\cdot A)A}{|A\times B|}\right|}{|B\times C|} = \frac{|(C-(C\cdot A)A)\times(B-(B\cdot A)A)|}{|A\times B|\,|A\times C|\,|B\times C|}$
In the denominator we get the product of the three vector products, namely $|A \times B|\,|B \times C|\,|C \times A|$, which is symmetric in the three vectors. Meanwhile in the numerator we get the norm of the vector

$$(C - (C \cdot A)A) \times (B - (B \cdot A)A). \tag{6.10.9}$$

The triple of vectors A, B, C is transformed into the triple A, $(C - (C \cdot A)A)$, $(B - (B \cdot A)A)$ by a volume-preserving transformation. This combined with the fact that A is a unit vector shows that the norm of the

[14] At a point where φ is not extremal as a function of θ, the theorem on the uniqueness of solution of ODE applies and gives a unique geodesic through the point. However, at a point of maximal φ, the hypothesis of the uniqueness theorem does not apply. Namely, the square root expression on the right hand side of (6.10.7) does not satisfy the Lipschitz condition as the expression under the square root sign vanishes. In fact, uniqueness fails at this point, as a latitude (which is not a geodesic) satisfies the differential equation, as well. Here we have $r = \text{const}$, and at an extremal value of φ one can no longer solve the equation by separation of variables (as this would involve division by the radical expression which vanishes at the extremal value of φ). At this point, there is a degeneracy and general results about uniqueness of solution cannot be applied.

vector (6.10.9) equals the absolute value of the determinant of the 3×3 matrix $[A\,B\,C]$ (i.e., absolute value of the volume of the parallelopiped spanned by the three vectors), which is also symmetric in the three points. Thus we have $\frac{\sin\alpha}{\sin a} = \frac{|\det[A\,B\,C]\,|}{\sin a \sin b \sin c}$, proving the spherical sine law. $\square$

THIRD PROOF OF SINE LAW. There is an alternative proof that relies on the identity $(a\times b)\times(a\times c) = (a\cdot(b\times c))a$ for each triple of vectors $a, b, c \in \mathbb{R}^3$. Indeed, $\sin\alpha = \frac{|(A\times B)\times(A\times C)|}{|A\times B|\,|A\times C|} = \frac{\det(A\,B\,C)}{\sin b \sin c}$ since A is a unit vector, and therefore $\frac{\sin\alpha}{\sin a} = \frac{\det(A\,B\,C)}{\sin a \sin b \sin c}$ as required. $\square$

The sources for the spherical sine law are Smart 1960, pp. 9-10; Gellert et al. 1989, p. 265; Zwillinger 1995, p. 469. Here (1) Gellert, W.; Gottwald, S.; Hellwich, M.; Kästner, H.; and Künstner, H. (Eds.). "Spherical Trigonometry." §12 in VNR Concise Encyclopedia of Mathematics, 2nd ed. New York: Van Nostrand Reinhold, pp. 261-282, 1989; (2) Smart, W. M. Text-Book on Spherical Astronomy, 6th ed. Cambridge, England: Cambridge University Press, 1960; (3) Zwillinger, D. (Ed.). "Spherical Geometry and Trigonometry." §6.4 in CRC Standard Mathematical Tables and Formulae. Boca Raton, FL: CRC Press, pp. 468-471, 1995.

CHAPTER 7

Geodesic equation

7.1. Gravitation, ants, and geodesics on a surface

The notion of a geodesic curve is important both in differential geometry and in relativity theory. What is a geodesic on a surface?

> A geodesic on a surface can be thought of as the path of an ant crawling along the surface of an apple,

according to the textbook *Gravitation* [**MiTW73**, p. 3].

Remark 7.1.1 (Narrow strip of apple's peel). Imagine that we peel off a narrow strip[1] of the apple's skin along the ant's trajectory, and then lay it out flat on a table. What we obtain is a straight line, revealing the ant's ability to travel along the shortest path.

On the other hand, a geodesic is defined by a certain nonlinear second order ordinary differential equation as in formula (7.2.1). To make the geodesic equation more concrete, we will examine the case of the surfaces of revolution. Here the geodesic equation transforms into a conservation law called conservation of angular momentum.[2]

In addition to a derivation in Section 7.2, we will also give a longer derivation of the geodesic equation using the calculus of variations.[3]

7.2. Geodesic equation on a surface

Consider the following data.

(1) A plane curve $\underset{s}{\mathbb{R}} \underset{\alpha}{\to} \underset{(u^1,u^2)}{\mathbb{R}^2}$ where $\alpha = (\alpha^1(s), \alpha^2(s))$.

(2) A parametrisation $\underline{x}\colon \mathbb{R}^2 \to \mathbb{R}^3$ of a surface M.

[1] retzua daka

[2] shimur tena' zaviti. This conservation law is equivalent to Clairaut's relation. The latter lends itself to a synthetic verification for spherical great circles; see Theorem 6.10.4. Thus Clairaut's relation provides a bridge between geometry and ODEs.

[3] I once heard R. Bott point out a surprising aspect of M. Morse's foundational work in this area. Namely, Morse systematically used the length functional on the space of curves. The simple idea of using the energy functional instead of the length functional was not exploited until later. The use of energy simplifies calculations considerably, as we will see in the optional Section 8.14.2.

(3) A curve $\beta = \underline{x} \circ \alpha$ on M given by the composition

$$\underset{s}{\mathbb{R}} \underset{\alpha}{\to} \underset{(u^1,u^2)}{\mathbb{R}^2} \underset{\underline{x}}{\to} \mathbb{R}^3.$$

We will first prove the following identity.

PROPOSITION 7.2.1. *Every regular curve $\beta(s)$ on M satisfies the identity*

$$\beta'' = \left(\alpha^{i\prime}\alpha^{j\prime}\Gamma_{ij}^k + \alpha^{k\prime\prime}\right) x_k + \left(L_{ij}\alpha^{j\prime}\alpha^{i\prime}\right) n$$

PROOF. Write $\beta = \underline{x} \circ \alpha$, then $\beta' = x_i(\alpha(s))\alpha^{i\prime}$ by chain rule. Differentiating again, we obtain

$$\begin{aligned}\beta'' &= \tfrac{d}{ds}(x_i \circ \alpha)\, \alpha^{i\prime} + x_i\, \alpha^{i\prime\prime} \\ &= x_{ij}\alpha^{j\prime}\alpha^{i\prime} + x_k\alpha^{k\prime\prime}.\end{aligned}$$

Meanwhile $x_{ij} = \Gamma_{ij}^k x_k + L_{ij} n$. Thus

$$\beta'' = \left(\Gamma_{ij}^k x_k + L_{ij} n\right) \alpha^{j\prime}\alpha^{i\prime} + x_k\alpha^{k\prime\prime}.$$

Rearranging the terms proves the proposition. □

DEFINITION 7.2.2. A *geodesic* is a curve $\beta = \underline{x} \circ \alpha$ on the surface M satisfying one of the following two equivalent conditions is satisfied:

(a) we have for each $k = 1, 2$,

$$(\alpha^k)'' + \Gamma_{ij}^k (\alpha^i)'(\alpha^j)' = 0 \quad \text{where} \quad {}' = \frac{d}{ds}\ , \tag{7.2.1}$$

meaning that

$$(\forall k) \quad \frac{d^2\alpha^k}{ds^2} + \Gamma_{ij}^k \frac{d\alpha^i}{ds}\frac{d\alpha^j}{ds} = 0; \tag{7.2.2}$$

(b) the vector β'' is perpendicular to the surface and one has

$$\beta'' = L_{ij}\alpha^{i\prime}\alpha^{j\prime} n. \tag{7.2.3}$$

REMARK 7.2.3. We will show in Lemma 8.1.1 that such a curve β must have constant speed.[4]

[4]In Section 8.14.2, the equations (7.2.1) will be derived using the calculus of variations.

7.3. Equivalence of definitions

In Section 7.2 we defined a geodesic curve $\beta(s) = x \circ \alpha(s)$ on a surface with parametrisation $x = x(u^1, u^2)$ in two ways that were claimed to be equivalent:

(a) we have for each $k = 1, 2$, $(\alpha^k)'' + \Gamma^k_{ij}(\alpha^i)'(\alpha^j)' = 0$;

(b) the vector β'' is normal to the surface and $\beta'' = L_{ij}\alpha^{i\prime}\alpha^{j\prime} n$.

Proof of equivalence. Assume β satisfies (a). We apply Proposition 7.2.1 to the effect that $\beta'' = \left(\alpha^{i\prime}\alpha^{j\prime}\Gamma^k_{ij} + \alpha^{k\prime\prime}\right) x_k + \left(L_{ij}\alpha^{j\prime}\alpha^{i\prime}\right) n$. Our assumption implies that the tangent component vanishes and we obtain

$$\beta'' = L_{ij}\alpha^{i\prime}\alpha^{j\prime} n,$$

showing that the vector β'' is normal to the surface.

Conversely, suppose β'' is proportional to the normal vector n. Then the tangential component of β'' must vanish, proving the equation $(\alpha^k)'' + \Gamma^k_{ij}(\alpha^i)'(\alpha^j)' = 0$ for each k. □

Corollary 7.3.1. *Every geodesic in the plane is a straight line.*

Proof. In the plane with a standard parametrisation, all coefficients vanish: $\Gamma^k_{ij} = 0$. Then the geodesic equation becomes $(\alpha^k)'' = 0$. Integrating twice, we obtain that $\alpha(s)$ is a linear function of s. Therefore the graph is a straight line, proving the corollary. □

CHAPTER 8

Geodesics, Weingarten map

8.1. Geodesic is constant speed

Let $\underline{x}(u^1, u^2)$ be a regular parametrisation of a surface M, and let $\beta = \underline{x} \circ \alpha$ be a smooth curve on M. Recall that β is a geodesic if it satisfies the geodesic equation

$$\alpha^{k''} + \Gamma^k_{ij}\alpha^{i'}\alpha^{j'} = 0 \tag{8.1.1}$$

for all k, and that if β is a geodesic then

$$\beta'' = L_{ij}\alpha^{i'}\alpha^{j'}n \tag{8.1.2}$$

(see equation (7.2.1)), where n is the normal vector to the surface at the point $\underline{x}(u^1, u^2)$.

Lemma 8.1.1. *On an arbitrary surface M, a curve β satisfying the geodesic equation* (8.1.1) *is necessarily constant speed.*

Proof. It suffices to prove that the square of the speed has vanishing derivative. From formula (8.1.2), we have

$$\begin{aligned}
\frac{d}{ds}\left(|\beta'|^2\right) &= 2\langle \beta'', \beta' \rangle \\
&= \langle L_{ij}\alpha^{i'}\alpha^{j'}n \,,\, \alpha^{k'}x_k \rangle \\
&= L_{ij}\alpha^{i'}\alpha^{j'}\alpha^{k'}\langle n, x_k \rangle \\
&= 0,
\end{aligned}$$

proving the lemma. □

8.2. Geodesics on a surface of revolution

Surfaces of revolution $M \subseteq \mathbb{R}^3$ are a rich source of interesting examples of surfaces. Because of the presence of a circle of symmetries, one can reduce the order of the differential equation of a geodesic from 2 to 1, making it easier to solve explicitly.

We use the notation $u^1 = \theta, u^2 = \phi$ for the parameters in the case of a surface of revolution generated by a plane curve $(r(\phi), z(\phi))$, and write $\underline{x}(\theta, \phi)$ in place of $\underline{x}(u^1, u^2)$. Here the function $r(\phi)$ is the distance from the point on the surface to the z-axis.

DEFINITION 8.2.1 (Latitude on a surface of revolution). On a surface of revolution M, the *latitude*[1] is the curve obtained by fixing $\phi = \phi_0$, and parametrized as follows:

$$\beta(\theta) = \underline{x}(\theta, \phi_0) = (r(\phi_0)\cos\theta, r(\phi_0)\sin\theta, z(\phi_0))$$

where $\theta \in [0, 2\pi]$.

LEMMA 8.2.2 (Angle with latitude). *Let M be a surface of revolution. Let $\beta(s) = \underline{x}(\theta(s), \phi(s))$ be a unit speed curve (not necessarily geodesic) on M. Then the angle $\gamma(s)$ at the point $\beta(s)$ between the curve and the latitude satisfies the relation*

$$\cos\gamma(s) = r\,\frac{d\theta}{ds},$$

where r is the distance from the point $\beta(s)$ to the z-axis.

PROOF. The tangent vector to the latitude is $x_1 = \frac{\partial x}{\partial\theta}$. Recall that $g_{11} = r^2(\phi)$ by Theorem 6.7.1, i.e., $|x_1| = r$; and $g_{12} = 0$. Then we can compute the cosine of the angle between two unit vectors by

$$\cos\gamma = \left\langle \frac{x_1}{|x_1|}, \frac{d\beta}{ds} \right\rangle.$$

Let $' = \frac{d}{ds}$. Recall that by chain rule $\beta' = x_1\,\theta' + x_2\,\phi'$. Thus

$$\begin{aligned}\cos\gamma &= \frac{1}{|x_1|}\langle x_1, \underbrace{x_1\theta' + x_2\phi'}_{\text{chain rule}}\rangle = \frac{\theta'}{|x_1|}\underbrace{\langle x_1, x_1\rangle}_{|x_1|^2} + \frac{\phi'}{|x_2|}\underbrace{\langle x_1, x_2\rangle}_{g_{12}=0}\\ &= \theta'|x_1| = r\,\theta',\end{aligned}$$

proving the lemma. □

8.3. Geodesic equation on a surface of revolution

We will use the shorthand notation $\theta(s), \phi(s)$ respectively for the components $\alpha^1(s), \alpha^2(s)$ of the curve $\alpha(s)$.

PROPOSITION 8.3.1. *On a surface of revolution, the differential equation of geodesic $\beta(s) = \underline{x}(\theta(s), \phi(s))$ for $k = 1$ becomes*

$$r\,\theta'' + 2\,\frac{dr}{d\phi}\,\theta'\phi' = 0. \tag{8.3.1}$$

[1] kav rochav

PROOF. The Gamma symbols for $k=1$, computed in Lemma 6.7.1, are $\Gamma_{11}^1 = \Gamma_{22}^1 = 0$ and $\Gamma_{12}^1 = \frac{\frac{dr}{d\phi}}{r}$. The differential equation of geodesic $\beta = x \circ \alpha$ for $k=1$ becomes

$$0 = \theta'' + 2\Gamma_{12}^1 \theta' \phi' = \theta'' + \frac{2\frac{dr}{d\phi}}{r}\theta' \phi',$$

as required. □

THEOREM 8.3.2. *On a surface of revolution $M \subseteq \mathbb{R}^3$ with coordinates (θ, ϕ), the differential equation of a geodesic for $k=1$ is equivalent to the first-order differential equation*

$$r^2\theta' = \text{const.} \tag{8.3.2}$$

PROOF. We multiply formula (8.3.1) by r to obtain

$$\begin{aligned} 0 &= r^2\theta'' + 2r\frac{dr}{d\phi}\theta'\phi' \\ &= (r^2\theta')' \end{aligned}$$

by Leibniz rule and chain rule. Integrating the equation $(r^2\theta')' = 0$ we obtain $r^2\theta' = \text{const}$ as required. □

COROLLARY 8.3.3. *The geodesic equation on a surface of revolution is equivalent to Clairaut's relation*

$$r\cos\gamma = \text{const}$$

along the geodesic, where γ is the angle with the latitude.[2]

PROOF. Since the geodesic β is constant speed by Lemma 8.1.1, we can assume the parameter s is arclength by rescaling the parameter by a constant factor. Now, by Lemma 8.2.2 we have $\cos\gamma = r\,\theta'$. Hence by formula (8.3.2) of Theorem 8.3.2, we have $r\cos\gamma = r^2\theta' = \text{const}$. Thus we obtain $r\cos\gamma = \text{const}$.[3] □

8.4. Geodesic equation in differential form

Recall that $(r(\phi), z(\phi))$ is the generating curve of a surface of revolution. We have $g_{11} = r^2(\phi)$ and $g_{22} = \left(\frac{dr}{d\phi}\right) + \left(\frac{dz}{d\phi}\right)^2$. Let $\beta = \underline{x} \circ \alpha(s)$ be a unit speed geodesic. Thus s is the arclength parameter and ds is an element of length.

[2]See Theorem 6.10.4.

[3]This generalizes Clairaut's relation from the sphere to arbitrary surfaces of revolution. In Section 6.10.5 we proved Clairaut's relation synthetically (using the spherical law of sines) for the sphere only. Corollary 8.3.3 shows that the relation holds for an arbitrary surface of revolution in $\mathbb{R}^3$.

Proposition 8.4.1. *On the surface of revolution generated by the curve $(r(\phi), z(\phi))$, every unit speed curve $\beta(s)$ satisfies the equation*

$$ds^2 = r^2 d\theta^2 + g_{22} d\phi^2,$$

or equivalently

$$\left(\frac{ds}{d\theta}\right)^2 = r^2 + g_{22}\left(\frac{d\phi}{d\theta}\right)^2.$$

Proof. We have $1 = \left\langle \frac{d\beta}{ds}, \frac{d\beta}{ds} \right\rangle = \langle x_1\theta' + x_2\phi', x_1\theta' + x_2\phi' \rangle = g_{11}(\theta')^2 + g_{22}(\phi')^2$. Thus,

$$1 = r^2\left(\frac{d\theta}{ds}\right)^2 + g_{22}\left(\frac{d\phi}{ds}\right)^2. \tag{8.4.1}$$

Multiplying (8.4.1) by $\left(\frac{ds}{d\theta}\right)^2$, we obtain $\left(\frac{ds}{d\theta}\right)^2 = r^2 + g_{22}\left(\frac{d\phi}{ds}\frac{ds}{d\theta}\right)^2 = r^2 + g_{22}\left(\frac{d\phi}{d\theta}\right)^2$, as required.[4] $\square$

See also note.[5]

[4] The geodesic equation for a surface of revolution can be solved explicitly in integrals, producing the following formula for θ as a function of ϕ: $\theta = c\int \frac{1}{r}\left(\frac{\left(\frac{dr}{d\phi}\right)^2 + \left(\frac{dz}{d\phi}\right)^2}{r^2 - c^2}\right)^{\frac{1}{2}} d\phi + c'$. Indeed, from equation (8.3.2) (equivalent to Clairaut's relation) we obtain $\frac{ds}{d\theta} = \frac{r^2}{c}$. By Proposition 8.4.1, we obtain the formula $\frac{r^4}{c^2} = r^2 + g_{22}\left(\frac{d\phi}{d\theta}\right)^2$ or, equivalently, $\frac{1}{c^2} = \frac{1}{r^2} + \frac{g_{22}}{r^4}\left(\frac{d\phi}{d\theta}\right)^2$ (which in the case of the sphere is the equation (6.10.5) of a great circle). Solving equation $\frac{r^4}{c^2} = r^2 + g_{22}\left(\frac{d\phi}{d\theta}\right)^2$ for $\frac{d\phi}{d\theta}$ we obtain $\frac{d\phi}{d\theta} = \left(\frac{\frac{r^4}{c^2} - r^2}{g_{22}}\right)^{1/2} = \frac{r}{c}\left(\frac{r^2 - c^2}{\left(\frac{dr}{d\phi}\right)^2 + \left(\frac{dz}{d\phi}\right)^2}\right)^{1/2}$. Solving for θ, we obtain $\theta = c\int \frac{1}{r}\sqrt{\frac{\left(\frac{dr}{d\phi}\right)^2 + \left(\frac{dz}{d\phi}\right)^2}{r^2 - c^2}}\, d\phi + c'$. As a corollary, we obtain the following: Along a great circle on S^2, we have the following relation between the spherical coordinates θ and ϕ: $\theta = c\int \frac{d\phi}{\sin\phi\sqrt{\sin^2\phi - c^2}} + c'$. Since $r(\phi) = \sin\phi$ and $z = \cos\phi$, we have obtained an equation of a great circle in spherical coordinates. Integrating this by a substitution $u = \cot\phi$ we obtain $\cot\phi = a\cos(\theta - \theta_0)$ where $a = \frac{1-c^2}{c}$. Note that the angles θ and ϕ are switched in the related discussion at `http://www.damtp.cam.ac.uk/user/reh10/lectures/nst-mmii-handout2.pdf`

[5] Area integration in polar coordinates: We review material from calculus on polar, cylindrical, and spherical coordinates as regards their role in integration. The polar coordinates (r, θ) in the plane arise naturally in complex analysis (of one complex variable). Area element in polar coordinates: Polar coordinates (koordinatot koteviot) (r, θ) satisfy $r^2 = x^2 + y^2$ and $x = r\cos\theta$, $y = r\sin\theta$. It is shown in elementary calculus that the area of a region $D \subseteq \mathbb{R}^2$ in the plane in polar coordinates is calculated using the following area element. The *area element* of polar coordinates is $dA = r\,dr\,d\theta$, meaning that the area of D relative to polar coordinates is computed as follows: $\int_D dA = \iint r dr d\theta$. Integration in cylindrical coordinates

8.5. Integration in spherical coordinates

Following some preliminaries on areas, directional derivatives, and Hessians, we will deal with a central object in the differential geometry of surfaces in Euclidean space, namely the Weingarten map, in Section 8.12.

Spherical coordinates[6] (ρ, θ, φ) in Euclidean 3-space are studied in Vector Calculus. They were already defined in Section 6.8.

volume element in spherical coordinates: The coordinate ρ is the distance from the point to the origin, satisfying $\rho^2 = x^2 + y^2 + z^2$, or $\rho^2 = r^2 + z^2$, where $r^2 = x^2 + y^2$. If we project the point orthogonally to the (x, y)-plane, the polar coordinates of its image, (r, θ), satisfy $x = r\cos\theta$ and $y = r\sin\theta$. The last coordinate φ of the spherical coordinates is the angle between the position vector of the point and the third basis vector e_3 in 3-space (pointing upward along the z-axis). Thus $z = \rho\cos\varphi$ while $r = \rho\sin\varphi$. Here we have the bounds $0 \leq \rho$, $0 \leq \theta \leq 2\pi$, and $0 \leq \varphi \leq \pi$ (note the different upper bounds for θ and φ).

Definition 8.5.1 (Volume element in spherical coordinates). The volume of a region in 3-space is calculated using the volume element of the form

$$dV = \rho^2 \sin\varphi \; d\rho \, d\varphi \; d\theta$$

meaning that the volume of a region $D \subseteq \mathbb{R}^3$ is

$$\int_D dV = \iiint_D \rho^2 \sin\varphi \; d\rho \, d\varphi \; d\theta.$$

Definition 8.5.2 (Area element on the sphere). The area of a spherical region $D \subseteq S^2$ on the unit sphere is calculated using an area element

$$\sin\varphi \; d\varphi \; d\theta.$$

See also the note.[7]

---

in $\mathbb{R}^3$. Cylindrical coordinates in Euclidean 3-space are studied in Vector Calculus. volume element (koordinatot gliliot) in cylindrical coordinates: (r, θ, z) are a natural extension of the polar coordinates (r, θ) in the plane. The volume of an open region $D \subseteq \mathbb{R}^3$ is calculated with respect to cylindrical coordinates using a suitable volume element. The volume element in cylindrical coordinates is $dV = r\,dr\,d\theta\,dz$, meaning that the volume of D is calculated as follows: $\int_D dV = \iiint r dr\,d\theta\,dz$.

[6]koordinatot kaduriot

[7]The area of a spherical region on a sphere of radius $\rho = \rho_1$ is calculated using the area element $dA = \rho_1^2 \sin\varphi \; d\varphi \; d\theta$. Thus the area of a spherical region D on a sphere of radius ρ_1 is given by the integral $\int_D dA = \iint \rho_1^2 \sin\varphi \; d\varphi \; d\theta = \rho_1^2 \iint \sin\varphi \; d\varphi \; d\theta$. An an example, calculate the area of the spherical region on

8.6. Measuring area on surfaces

Let $\underline{x}(u^1, u^2)$ be a parametrisation of a surface $M \subseteq \mathbb{R}^3$. The area of the parallelogram spanned by the tangent vectors x_1 and x_2 can be calculated as the square root of their Gram matrix, namely the matrix of the first fundamental form. This motivates the following definition.

Definition 8.6.1 (Area element of a surface). The *area* of the surface M parametrized by $\underline{x}\colon U \to \mathbb{R}^3$ is computed by integrating the area element

$$\sqrt{g_{11}g_{22} - g_{12}^2}\, du^1 du^2 \tag{8.6.1}$$

over the domain U of the map x. Thus

$$\mathrm{area}(M) = \int_U \sqrt{\det(g_{ij})}\, du^1 du^2,$$

where M is the region parametrized by $\underline{x}(u^1, u^2)$.

Remark 8.6.2. The presence of the square root in the formula is explained in infinitesimal calculus in terms of the Gram matrix, *cf.* formula (5.5.1). The geometric meaning of the square root is the area of the parallelogram spanned by the pair of standard (coordinate) tangent vectors.

Example 8.6.3. Consider the parametrisation of the unit sphere provided by spherical coordinates. Then the area element is[8]

$$\sin\varphi\ d\varphi\ d\theta$$

where $\sin\varphi = \sqrt{\det(g_{ij})}$. We recover the formula familiar from calculus for the area of a region D on the unit sphere:

$$\mathrm{area}(D) = \iint_D \sin\varphi\ d\varphi\ d\theta$$

as in Example 8.5.2.

8.7. Directional derivative as derivative along path

We will represent a vector $v \in \mathbb{R}^n$ as the velocity vector

$$v = \tfrac{d\alpha}{dt} = \dot\alpha(0)$$

of a curve $\alpha(t)$, at $t = 0$. Typically we will be interested in the case $n = 3$ (or 2).

---

a sphere of radius ρ_1 contained in the first octant (so that all three Cartesian coordinates are positive).

[8]Here there is a nasty problem of sign, since we chose $u^1 = \theta$. At the level of calculus this does not matter but it becomes important when integrating differential 2-forms.

Definition 8.7.1 (Directional derivative along curve). Given a function f of n variables, its directional derivative[9] $\nabla_v f$ at a point $p \in \mathbb{R}^n$, in the direction of a vector v is defined by setting

$$\nabla_v f = \left.\frac{d\,(f \circ \alpha(t))}{dt}\right|_{t=0}$$

where $\alpha(0) = p$ and $\dot{\alpha}(0) = v$.

Lemma 8.7.2. *The value of the directional derivative is independent of the choice of the curve $\alpha(t)$ representing the vector v.*

Proof. The lemma is proved in *Elementary Calculus* [**Ke74**]. $\square$

Let $p = \underline{x}(u^1, u^2)$ be a point of a surface $M \subseteq \mathbb{R}^3$. The tangent plane to the surface at the point p, denoted T_p, was defined in Section 5.5 and is the plane passing through p and spanned by vectors $\underline{x}_1$ and x_2:

$$T_p = \mathrm{Span}_{\mathbb{R}}(x_1, x_2).$$

It can be thought of as the plane perpendicular to the normal vector n at p.

Definition 8.7.3. Denote by $\mathrm{Span}_{\mathbb{R}} n$ the line spanned by n.

Definition 8.7.4 (Orthogonal decomposition). We have an orthogonal decomposition

$$\mathbb{R}^3 = T_p + \mathrm{Span}_{\mathbb{R}} n,$$

meaning that each vector $v \in T_p$ is orthogonal to n.

Example 8.7.5. Suppose $\underline{x}(u^1, u^2)$ is a parametrisation of the unit sphere $S^2 \subseteq \mathbb{R}^3$. At a point $p = (a, b, c) \in S^2$, the normal vector is the position vector itself:

$$n = \begin{pmatrix} a \\ b \\ c \end{pmatrix} = p.$$

In other words, $n(u^1, u^2) = \underline{x}(u^1, u^2) = p$ and we have an orthogonal decomposition $\mathbb{R}^3 = T_p + \mathrm{Span}_{\mathbb{R}}\, p$.

Remark 8.7.6. The sphere is defined implicitly by $F(x, y, z) = 0$ where $F(x, y, z) = x^2 + y^2 + z^2 - 1$. We have $\nabla F = (2x, 2y, 2z)$ and therefore the gradient ∇F at a point is proportional to the radius vector of the point on the sphere (see Lemma 8.8.1 below).

[9]nigzeret kivunit

8.8. Extending v. field along surface to an open set in $\mathbb{R}^3$

Recall the following from elementary calculus.

LEMMA 8.8.1. *The gradient ∇F of a function $F = F(x, y, z)$ at a point where the gradient is nonzero, is perpendicular to the level surface $\{(x, y, z) \colon F(x, y, z) = 0\}$ of the function F.*

Consider a regular parametrisation $\underline{x}(u^1, u^2)$ of a neighborhood of a point $p \in M \subseteq \mathbb{R}^3$ and its normal vector $n = n(u^1, u^2)$. In the set-up $\underset{t}{\mathbb{R}} \underset{\alpha}{\longrightarrow} \underset{(u^1,u^2)}{\mathbb{R}^2} \underset{\underline{x}}{\longrightarrow} \mathbb{R}^3$, consider the curve $\beta = \underline{x} \circ \alpha$. Let $p = \beta(0)$. Let $v = \frac{d\beta}{dt}\big|_{t=0}$ be the initial vector of the curve $\beta(t)$. By chain rule, we have

$$v = \frac{d\alpha^i}{dt} x_i.$$

Now consider the normal vector to the surface M

$$n \circ \alpha(t)$$

varying along the curve $\beta(t)$ on M. This vector is only defined along the curve. It is not defined in any open neighborhood in $\mathbb{R}^3$. We would like to extend it to a vector field in an open neighborhood of the point $p \in \mathbb{R}^3$, so as to be able to differentiate it.[10]

THEOREM 8.8.2. *The normal vector n can be extended to a differentiable vector field $N(x, y, z)$ defined in an open neighborhood of $p \in \mathbb{R}^3$, in the sense that we have*

$$n(u^1, u^2) = N\left(\underline{x}(u^1, u^2)\right). \tag{8.8.1}$$

PROOF. In a neighborhood of p, we can represent the surface M implicitly as the level surface $F(x, y, z) = 0$, where $\nabla F \neq 0$ at p. By Lemma 8.8.1, the normalisation $N = \frac{1}{|\nabla F|}\nabla F$ of the gradient ∇F of F gives the required extension in the sense of (8.8.1) of the normal n. □

COROLLARY 8.8.3. *Since the vector field N is defined in an open neighborhood of a point, the usual notion of directional derivative $\nabla_v N$ can be applied to N.*

8.9. Differentiating n and N

Note the following six points.

(1) $M \subseteq \mathbb{R}^3$ is a surface;
(2) $\underline{x}(u^1, u^2)$ is a regular parametrisation of M;

[10] The notion of vector field is defined in infi 4 but only the maslul of applied mathematics requires it. The maslul of pure mathematics takes a course on rings instead.

(3) $n(u^1, u^2)$ is the normal vector field along M;
(4) $N(x, y, z)$ is a vector field defined in an open neighborhood in $\mathbb{R}^3$, and extends $n(u^1, u^2)$ in the sense that

$$n(u^1, u^2) = N(\underline{x}(u^1, u^2)), \tag{8.9.1}$$

as shown in Section 8.8;
(5) ∇_v is the directional derivative defined in Section 8.7 via a representative curve for v;
(6) N is a vector-valued function defined in an open neighborhood in $\mathbb{R}^3$, and therefore can be differentiated.

PROPOSITION 8.9.1. *Let $p \in M$, and $v \in T_pM$ where $v = \beta'(0)$ and $\beta(t) = \underline{x}(\alpha(t))$. Then the directional derivative $\nabla_v N$ satisfies*

$$\nabla_v N = \left.\frac{d\,(n \circ \alpha(t))}{dt}\right|_{t=0}$$

at the point p.

PROOF. By (8.9.1), the function $n(\alpha(t))$ satisfies

$$n(\alpha(t)) = N(\beta(t)), \tag{8.9.2}$$

where $\beta = \underline{x} \circ \alpha$. By Lemma 8.7.2, the directional derivative is independent of the choice of the representing curve. Hence the directional derivative $\nabla_v N$ can be calculated using the particular curve β. By (8.9.2), we obtain

$$\nabla_v N = \frac{d\,(N \circ \beta(t))}{dt} = \frac{d\,(n \circ \alpha(t))}{dt},$$

as required. □

We will exploit this proposition to define the Weingarten map in Section 8.12, after motivating it in terms of the Hessian in Section 8.10.

8.10. Hessian of a function at a critical point

This section is intended to motivate the definition of the Weingarten map in Section 8.12. The key observation is Remark 8.11 below, relating the Hessian and the Weingarten map.

PROPOSITION 8.10.1. *Let $f(x, y)$ of two variables. Let p be a critical point p of f. Let $M \subseteq \mathbb{R}^3$ be the graph of f. Then the tangent plane of M at the point $(p, f(p)) \in M$ is a horizontal plane.*

PROOF. By Theorem 6.2.3, the unit normal vector n to the surface M is proportional to $(f_x, f_y, -1)^t$ up to sign. At the critical point, we have $\nabla f = 0$ so that $f_x = f_y = 0$ and therefore $n = \pm(0, 0, 1)^t = e_3$. □

The Hessian (matrix of second derivatives) of the function at the critical point captures the main features of the local behavior of the function in a neighborhood of the critical point up to negligible higher order terms. Thus, we have the following typical result concerning the surface given by the graph of the function in $\mathbb{R}^3$.

THEOREM 8.10.2 (Sign of Hessian). *Let p be a critical point of f. If $\det H_f < 0$ then p is a saddle point. If $\det H_f > 0$ then p is a local minimum or maximum.*

8.11. Hessian as linear map

Consider the graph M of f at a critical point p of f. Consider the Hessian matrix $H_f = (f_{ij})$. We now think of the matrix H_f as defining an endomorphism of the horizontal plane $T_p = \mathbb{R}^2$:

$$H_f : T_p \to T_p$$

by the usual formula $w = H_f(v)$ if and only if

$$w_i = \sum_{j=1}^{2} f_{ij} v_j .$$

Then the Hessian of a function at a critical point becomes a special case of the *Weingarten* map, defined in Section 8.12. The determinant of the Weingarten map plays a special role in determining the geometry of the surface near p, similar to that of the Hessian noted in Corollary 8.10.2.

8.12. Weingarten map as coordinate-free Hessian

The definition of the Weingarten operator can be viewed as an analogue of a formula involving the derivative of the normal vector that we saw in the context of curves. Recall the following.

PROPOSITION 8.12.1 (Comparison with curves). *For a plane curve $\alpha(s)$ with tangent vector $v(s)$ and normal vector $n(s)$, we have the equation*

$$\frac{d}{ds} n(s) = \pm k_\alpha(s) v(s). \tag{8.12.1}$$

This was proved in Proposition 4.3.7.

COROLLARY 8.12.2. *The curvature is the rate of change of the normal vector:*

$$k_\alpha(s) = \left| \frac{dn}{ds} \right| .$$

The Weingarten map W_p at a point $p \in M$ can be thought of as a surface analog of the formula (8.12.1) for curves. Namely W_p carries the information about the curvature of M.

In Section 8.10, we considered the special case of a surface given by the graph of a function f of two variables near a critical point of f. Now consider the more general framework of a parametrized regular surface M in $\mathbb{R}^3$ with a regular parametrisation $\underline{x}(u^1, u^2)$.

Instead of working with a matrix of second derivatives, we will give a definition of an endomorphism of the tangent plane. This is a coordinate-free way of talking about the Hessian matrix.

Theorem 8.8.2 enables us to extend the normal vector field n along the surface to a vector field $N(x, y, z)$ defined in an open neighborhood of $p \in \mathbb{R}^3$, so that $n(u^1, u^2) = N(\underline{x}(u^1, u^2))$.

Definition 8.12.3 (Weingarten map as a coordinate-free way of talking about the Hessian). Let $p \in M$. Let $T_p = T_pM$ be the tangent plane to the surface at p. The *Weingarten map* (also known as the shape operator)

$$W_p \colon T_p \to T_p$$

is the endomorphism of the tangent plane given by the directional derivative of the vector-valued function N:

$$W_p(v) = \nabla_v N = \left.\frac{d}{dt}\right|_{t=0} n \circ \alpha(t), \tag{8.12.2}$$

where the curve $\beta = \underline{x} \circ \alpha$ is chosen so that $\beta(0) = p$ while $\beta'(0) = v$.

We will show that W_p is a well-defined map in Section 8.13.

8.13. Properties of Weingarten map

Consider a surface $M \subseteq \mathbb{R}^3$. The Weingarten map

$$W_p \colon T_p \to T_p$$

at a point $p \in M$ was defined in Section 8.12. The following points should be kept in mind:

(1) Suppose p is a critical point of a function f and the surface M is the graph of f. Then the Weingarten map of M at p with respect to the standard coordinates is expressed by the Hessian matrix of f:

$$H_f = \left(f_{ij}\right)_{\substack{i=1,2\\ j=1,2}}$$

(see Theorem 9.1.1 for details).

(2) The Weingarten map at p measures the way the unit normal vector to the surface changes as the point $p \in M$ varies.

LEMMA 8.13.1. *The map W_p is well defined in that the image is included in the tangent plane T_p at p.*

PROOF. Let $v \in T_p$ be a tangent vector. We choose a curve $\beta(t) = \underline{x} \circ \alpha(t)$ so that $v = \frac{d\beta}{dt}$ when $t = 0$. The right-hand side of

$$W_p(v) = \nabla_v N = \left.\frac{d}{dt}\right|_{t=0} n \circ \alpha(t), \tag{8.13.1}$$

as in formula (8.12.2) is a priori a vector in $\mathbb{R}^3$. We have to show that it is indeed produces a vector that lies in the tangent plane T_p i.e., is perpendicular to the unit normal vector n. Using (8.13.1) and applying the Leibniz rule, we obtain

$$\begin{aligned}\langle W_p(v), n(\alpha(t))\rangle &= \langle \nabla_v N, n(\alpha(t))\rangle \\ &= \frac{1}{2}\frac{d}{dt}\langle n \circ \alpha(t), n \circ \alpha(t)\rangle \\ &= \frac{1}{2}\frac{d}{dt}(1) \\ &= 0\end{aligned}$$

since n is a unit vector at every point of M. Therefore the image vector $W_p(v)$ is orthogonal to n. Hence it lies in T_p, as required. $\square$

8.14. Self-adjointness of Weingarten map

Let $p \in M$. Recall that the Weingarten map $W_p \colon T_p \to T_p$ is defined as follows. Let $v \in T_p$. Then $W_p(v) = \nabla_v N$ where $N = N(x, y, z)$ is the extension to an open neighborhood of $p \in \mathbb{R}^3$ of the unit normal vector $n = n(u^1, u^2)$ along M.

The connection with the matrix of second derivatives of the parametrisation $\underline{x}$ is given in the following theorem.

THEOREM 8.14.1. *The map W_p has the following properties:*

(1) *it is a selfadjoint endomorphism of the tangent plane T_p, and*
(2) *it satisfies $\forall i, j,\ \langle W_p(x_i), x_j\rangle = -\left\langle n, \frac{\partial^2 \underline{x}}{\partial u^i \partial u^j}\right\rangle$.*

PROOF. By definition of the Weingarten map, $\nabla_v N$ is the derivative of N along a curve with initial vector $v \in T_p$. To simplify notation, assume that $p = \underline{x}(0, 0)$.

Step 1. We can choose the particular curve $\gamma(t) = \underline{x}(t, 0)$ varying along the first coordinate. Then we have $\frac{d\gamma}{dt} = x_1$ by definition of partial derivatives. Therefore

$$\nabla_{x_1} N = \frac{d}{dt} N(\gamma(t)) = \frac{\partial n}{\partial u^1}$$

by definition of the directional derivative.

Step 2. Using the curve $\delta(t) = \underline{x}(0,t)$ varying along the second coordinate, we similarly obtain $\nabla_{x_2} N = \frac{\partial n}{\partial u^2}$. Thus

$$\forall j = 1, 2, \quad \nabla_{x_j} N = \frac{\partial n}{\partial u^j}.$$

Step 3. We have by Leibniz rule

$$\begin{aligned} \langle W_p(x_i), x_j \rangle &= \left\langle \frac{\partial n}{\partial u^i}, x_j \right\rangle \\ &= \frac{\partial}{\partial u^i} \langle n, x_j \rangle - \langle n, \frac{\partial}{\partial u^i} x_j \rangle \\ &= -\left\langle n, \frac{\partial^2 x}{\partial u^i \partial u^j} \right\rangle. \end{aligned}$$

Thus we obtain that for all i, j,

$$\langle W_p(x_i), x_j \rangle = -\left\langle n, \frac{\partial^2 x}{\partial u^i \partial u^j} \right\rangle. \tag{8.14.1}$$

Step 4. The right-hand side of equation (8.14.1) is an expression symmetric in i and j by equality of mixed partials. Thus,

$$(\forall i, j) \quad \langle W_p(x_i), x_j \rangle = \langle x_i, W_p(x_j) \rangle. \tag{8.14.2}$$

This proves the selfadjointness of W_p by verifying it for a set of basis vectors. □

COROLLARY 8.14.2. *The eigenvalues of the Weingarten map are real.*

PROOF. The endomorphism W_p of the vector space T_pM is selfadjoint and we apply Corollary 2.3.2. □

DEFINITION 8.14.3. The *principal curvatures* at $p \in M$, denoted k_1 and k_2, are the eigenvalues of the Weingarten map W_p.

By Corollary 8.14.2, the principal curvatures are real. The principal curvatures will be discussed in more detail in Section 9.10.

8.14.1. Normal and geodesic curvatures. The material in this section is optional. Let $\underline{x}(u^1, u^2)$ be a surface in $\mathbb{R}^3$. Let $\alpha(s) = (\alpha^1(s), \alpha^2(s))$ a curve in $\mathbb{R}^2$, and consider the curve $\beta = x \circ \alpha$ on the surface $\underline{x}$. Consider also the normal vector to the surface, denoted n. The tangent unit vector $\beta' = \frac{d\beta}{ds}$ is perpendicular to n, so $\beta', n, \beta' \times n$ are three unit vector spanning $\mathbb{R}^3$. As β' is a unit vector, it is perpendicular to β'', and therefore β'' is a linear combination of n and $\beta' \times n$. Thus $\beta'' = k_n n + k_g(\beta' \times n)$, where $k_n, k_g \in \mathbb{R}$ are called the *normal* and the *geodesic curvature* (resp.). Note that $k_n = \beta'' \cdot n,\ k_g = \beta'' \cdot (\beta' \times n)$. If k is the curvature of β, then we have that $k^2 = ||\beta''||^2 = k_n^2 + k_g^2$.

EXERCISE 8.14.4. *Let $\beta(s)$ be a unit speed curve on a sphere of radius r. Then the normal curvature of β is $\pm 1/r$.*

REMARK 8.14.5. *Let π be a plane passing through the center of the sphere S in Exercise 8.14.4, and Let $C = \pi \cap S$. Then the curvature k of C is $1/r$, and thus $k_g = 0$. On the other hand, if π does not pass through the center of S (and $\pi \cap S \neq \emptyset$) then the geodesic curvature of the intersection $\neq 0$.*

PROPOSITION 8.14.6. *If a unit speed curve β on a surface is geodesic then its geodesic curvature is $k_g = 0$.*

PROOF. If β is a geodesic, then β'' is parallel to the normal vector n, so it is perpendicular to $n \times \beta''$, and therefore $k_g = \beta'' \cdot (n \times \beta'') = 0$. □

EXERCISE 8.14.7. *The inverse direction of Proposition 8.14.6 is also correct. Prove it.*

EXAMPLE 8.14.8. *Any line $\gamma(t) = at + b$ on a surface is a geodesic, as $\gamma'' = 0$ and therefore $k_g = 0$.*

EXAMPLE 8.14.9. *Take a cylinder x of radius 1 and intersect it with a plane π parallel to the xy plane. Let $C = x \cap \pi$ - it is a circle of radius 1. Thus $k = 1$. Show that $k_n = 1$ and thus C is a geodesic.*

The principal curvatures, denoted k_1 and k_2, are the eigenvalues of the Weingarten map W. Assume $k_1 > k_2$.

THEOREM 8.14.10. *The minimal and maximal values of the normal curvature k_n at a point p of all curves on a surface passing through p are k_2 and k_1.*

PROOF. This is proven using the fact that $k_g = 0$ for geodesics. □

8.14.2. Calculus of variations and the geodesic equation. The material in this subsection is optional. Calculus of variations is known as tachsiv variatsiot. Let $\alpha(s) = (\alpha^1(s), \alpha^2(s))$, and consider the curve $\beta = x \circ \alpha$ on the surface given by $[a,b]_s \to^{\alpha} \mathbb{R}^2 \to^{x} \mathbb{R}^3$. Consider the energy functional $\mathcal{E}(\beta) = \int_a^b \|\beta'\|^2 ds$. Here $' = \frac{d}{ds}$. We have $\frac{d\beta}{ds} = x_i (\alpha^i)'$ by chain rule. Thus

$$\mathcal{E}(\beta) = \int_a^b \langle \beta', \beta' \rangle ds = \int_a^b \left\langle x_i \alpha^{i'}, x_j \alpha^{j'} \right\rangle ds = \int_a^b g_{ij} \alpha^{i'} \alpha^{j'} ds. \quad (8.14.3)$$

Consider a variation $\alpha(s) \to \alpha(s) + t\delta(s)$, t small, such that $\delta(a) = \delta(b) = 0$. If $\beta = x \circ \alpha$ is a critical point of $\mathcal{E}$, then for any perturbation δ vanishing at the endpoints, the following derivative vanishes:

$$\begin{aligned}
0 &= \frac{d}{dt}\bigg|_{t=0} \mathcal{E}(x\circ(\alpha+t\delta)) \\
&= \frac{d}{dt}\bigg|_{t=0} \left\{\int_a^b g_{ij}(\alpha^i+t\delta^i)'(\alpha^j+t\delta^j)'\,ds\right\} \text{(from equation (8.14.3))} \\
&= \underbrace{\int_a^b \frac{\partial(g_{ij}\circ(\alpha+t\delta))}{\partial t}\alpha^{i'}\alpha^{j'}\,ds}_{A} + \underbrace{\int_a^b g_{ij}\big(\alpha^{i'}\delta^{j'}+\alpha^{j'}\delta^{i'}\big)ds}_{B},
\end{aligned}$$

so that we have

$$A+B=0. \tag{8.14.4}$$

We will need to compute both the t-derivative and the s-derivative of the first fundamental form. The formula is given in the lemma below.

Lemma 8.14.11. *The partial derivatives of* $g_{ij} = g_{ij}\circ(\alpha(s)+t\delta(s))$ *along* $\beta = x\circ\alpha$ *are given by the following formulas:* $\frac{\partial}{\partial t}(g_{ij}\circ(\alpha+t\delta)) = (\langle x_{ik},x_j\rangle+\langle x_i,x_{jk}\rangle)\delta^k$, *and* $\frac{\partial g_{ik}}{\partial s} = (\langle x_{im},x_k\rangle+\langle x_i,x_{km}\rangle)(\alpha^m)'$.

Proof. We have

$$\begin{aligned}
\frac{\partial}{\partial t}(g_{ij}\circ(\alpha+t\delta)) &= \frac{\partial}{\partial t}\langle x_i\circ(\alpha+t\delta)), x_j\circ(\alpha+t\delta)\rangle \\
&= \left\langle \frac{\partial}{\partial t}(x_i\circ(\alpha+t\delta)), x_j\right\rangle + \left\langle x_i, \frac{\partial}{\partial t}(x_j\circ(\alpha+t\delta))\right\rangle \\
&= \langle x_{ik}\delta^k, x_j\rangle + \langle x_i, x_{jk}\delta^k\rangle = (\langle x_{ik},x_j\rangle+\langle x_i,x_{jk}\rangle)\delta^k.
\end{aligned}$$

Furthermore, $g'_{ik} = \langle x_i,x_k\rangle' = \langle x_{im}(\alpha^m)', x_k\rangle + \langle x_i, x_{km}(\alpha^m)'\rangle$. □

Lemma 8.14.12. *Let* $f\in C^0[a,b]$. *Suppose that for all* $g\in C^0[a,b]$ *we have* $\int_a^b f(x)g(x)dx = 0$. *Then* $f(x)\equiv 0$. *This conclusion remains true if we use only test functions* $g(x)$ *such that* $g(a)=g(b)=0$.

Proof. We try the test function $g(x)=f(x)$. Then $\int_a^b (f(x))^2 ds = 0$. Since $(f(x))^2 \geq 0$ and f is continuous, it follows that $f(x)\equiv 0$. If we want $g(x)$ to be 0 at the endpoints, it suffices to choose $g(x) = (x-a)(b-x)f(x)$. □

Theorem 8.14.13. *Suppose* $\beta = x\circ\alpha$ *is a critical point of the energy functional (endpoints fixed). Then* β *satisfies the differential equation* $(\forall k) \quad (\alpha^k)'' + \Gamma^k_{ij}(\alpha^i)'(\alpha^j)' = 0$.

We use Lemma 8.14.11 to evaluate the term A from equation (8.14.4) as follows: $A = \int_a^b(\langle x_{ik},x_j\rangle+\langle x_i,x_{jk}\rangle)(\alpha^i)'(\alpha^j)'\delta^k ds = 2\displaystyle\int_a^b \langle x_{ik},x_j\rangle\alpha^{i'}\alpha^{j'}\delta^k ds$, since summation is over both i and j. Similarly, $B = 2\int_a^b g_{ij}\alpha^{i'}\delta^{j'}ds =$

$-2\int_a^b \left(g_{ij}\alpha^{i'}\right)' \delta^j ds$ by integration by parts, where the boundary term vanishes since $\delta(a) = \delta(b) = 0$. Hence $B = -2\int_a^b \left(g_{ik}\alpha^{i'}\right)' \delta^k ds$ by changing an index of summation. Thus $\frac{1}{2}\frac{d}{dt}\Big|_{t=0}(\mathcal{E}) = \int_a^b \left\{\langle x_{ik}, x_j\rangle \alpha^{i'}\alpha^{j'} - \left(g_{ik}\alpha^{i'}\right)'\right\} \delta^k ds.$ Since this is true for any variation (δ^k), by Lemma 8.14.12 we obtain the Euler-Lagrange equation $(\forall k)\quad \langle x_{ik}, x_j\rangle \alpha^{i'}\alpha^{j'} - \left(g_{ik}\alpha^{i'}\right)' \equiv 0$, or

$$\langle x_{ik}, x_j\rangle \alpha^{i'}\alpha^{j'} - {g_{ik}}'\alpha^{i'} - g_{ik}\alpha^{i''} = 0. \tag{8.14.5}$$

Using the formula from Lemma 8.14.11 for the s-derivative of the first fundamental form, we can rewrite the formula (8.14.5) as follows:

$$\begin{aligned}
0 &= \langle x_{ik}, x_j\rangle \alpha^{i'}\alpha^{j'} - \langle x_{im}, x_k\rangle \alpha^{m'}\alpha^{i'} - \langle x_i, x_{km}\rangle \alpha^{i'}\alpha^{m'} - g_{ik}\alpha^{i''} \\
&= -\langle \Gamma_{im}^n x_n, x_k\rangle \alpha^{m'}\alpha^{i'} - g_{ik}\alpha^{i''} \\
&= -\Gamma_{im}^n g_{nk}\alpha^{m'}\alpha^{i'} - g_{ik}\alpha^{i''},
\end{aligned}$$

where the cancellation of the first and the third term in the first line results from replacing index i by j and m by i in the third term. This is true $\forall k$. Now multiply by g^{jk} to obtain $g^{jk}\Gamma_{im}^n g_{nk}\alpha^{m'}\alpha^{i'} + g^{jk}g_{ik}\alpha^{i'} = \delta_n^j\Gamma_{im}^n\alpha^{m'}\alpha^{i'} + \delta_i^j\alpha^{i''} = \Gamma_{im}^j\alpha^{m'}\alpha^{i'} + \alpha^{j''} = 0$, which is the desired geodesic equation.

CHAPTER 9

Gaussian curvature; second fundamental form

9.1. Relation to the Hessian of a function

Let $p \in M \subseteq \mathbb{R}^3$. The Weingarten map $W_p\colon T_p \to T_p$ can be understood as a generalisation of the Hessian matrix in the following sense. Recall from (8.14.1) that

$$\langle W_p(x_i), x_j \rangle = -\left\langle n, \frac{\partial^2 \underline{x}}{\partial u^i \partial u^j} \right\rangle. \tag{9.1.1}$$

Theorem 9.1.1 (Relation to the Hessian). *Given a function f of two variables, consider its graph $M \subseteq \mathbb{R}^3$ with parametrisation $\underline{x}(u^1, u^2) = (u^1, u^2, f(u^1, u^2))$. Let p be a critical point of f and let $n = -e_3$. Then the inner products $\langle W_p(x_i), x_j\rangle$ at p form the Hessian matrix*

$$H_f(p) = \left(\frac{\partial^2 f}{\partial u^i \partial u^j} \right)$$

of f at $p = \underline{x}(u^1, u^2)$.

Proof. The second partial derivatives of the parametrisation are $x_{ij} = (0, 0, f_{ij})^t$. The normal vector at a critical point is $n = -e_3$. Therefore we obtain $\langle x_{ij}, n\rangle = -f_{ij}$. The theorem now follows from formula (9.1.1). $\square$

Example 9.1.2. Consider the plane with its standard parametrisation $\underline{x}(u^1, u^2) = (u^1, u^2, 0)$. We have $x_1 = e_1$, $x_2 = e_2$, while the normal vector field n

$$n(u^1, u^2) = \frac{x_1 \times x_2}{(x_1 \times x_2)} = e_1 \times e_2 = e_3$$

is constant. We can therefore extend it to a constant vector field N defined on all of $\mathbb{R}^3$. Thus $\nabla_v N \equiv 0$ and $W_p(v) \equiv 0$, and the Weingarten map is identically zero. This can be seen also from the vanishing of the second derivatives of the parametrisation.

In Sections 9.2 and 9.3 we will present nonzero examples of the Weingarten map.

9.2. Weingarten map of sphere

We defined the Weingarten map $W_p \colon T_p \to T_p$ of a surface $M \subseteq \mathbb{R}^3$ in Section 8.12.

THEOREM 9.2.1. *Let $M \subseteq \mathbb{R}^3$ be the sphere of radius $R > 0$. Then the Weingarten map at every point $p \in M$ of the sphere is the scalar map $W_p \colon T_p \to T_p$ given by*

$$W_p = \frac{1}{R}\mathrm{Id}_T$$

where Id_T is the identity map of the tangent plane T_p.

PROOF. Represent a tangent vector $v \in T_p$ by $v = \beta'(0)$ where $\beta(t) = \underline{x} \circ \alpha(t)$ as usual. On the sphere M of radius $R > 0$, we have $n(\alpha(t)) = \frac{1}{R}\beta(t)$ (normal vector is the normalized position vector). Hence

$$\begin{aligned} W_p(v) &= \nabla_v N(\beta(t)) \\ &= \frac{d}{dt}\Big|_{t=0} n \circ \alpha(t) \\ &= \frac{1}{R}\frac{d}{dt}\Big|_{t=0} \beta(t) \\ &= \frac{1}{R} v. \end{aligned}$$

Thus $W_p(v) = \frac{1}{R} v$ for all $v \in T_p$. In other words, $W_p = \frac{1}{R}\,\mathrm{Id}_T$. □

Note that the Weingarten map has rank 2 in this case, i.e., it is invertible.

9.3. Weingarten map of cylinder

The cylinder is parametrized by $\underline{x}(u^1, u^2) = (\cos u^1, \sin u^1, u^2)$, with

$$\begin{cases} x_1 = (-\sin u^1, \cos u^1, 0)^t \\ x_2 = e_3, \\ n = (\cos u^1, \sin u^1, 0)^t. \end{cases}$$

THEOREM 9.3.1. *For the cylinder we have $W_p(x_1 + cx_2) = x_1$ for all $c \in \mathbb{R}$.*

PROOF. Extend n to a vector field N defined in an open neighborhood in $\mathbb{R}^3$, with the usual relation $\nabla_{x_i} N = \frac{\partial}{\partial u^i} n$. Hence we have

$$\nabla_{x_1} N = \frac{\partial}{\partial u^1} n = (-\sin u^1, \cos u^1, 0)^t. \tag{9.3.1}$$

Differentiating with respect to u_2, we obtain

$$\nabla_{x_2} N = \frac{\partial}{\partial u^2} n = (0,0,0)^t. \tag{9.3.2}$$

Now let $v = v^1 x_1 + v^2 x_2$ be an arbitrary tangent vector. By linearity of directional derivatives, we obtain

$$\nabla_v N = \nabla_{v^i x_i} N = v^1 \nabla_{x_1} N + v^2 \nabla_{x_2} N$$

Hence (9.3.1) and (9.3.2) yield

$$\nabla_v N = v^1 \nabla_{x_1} N = v^1 \begin{pmatrix} -\sin u^1 \\ \cos u^1 \\ 0 \end{pmatrix},$$

and therefore,

$$W_p(v) = v^1 \begin{pmatrix} -\sin u^1 \\ \cos u^1 \\ 0 \end{pmatrix} = v^1 x_1,$$

where $v = v^i x_i$. Thus, $W_p(x_1 + cx_2) = x_1$ for all $c \in \mathbb{R}$. □

Note that the Weingarten map has rank 1 in this case.

9.4. Coefficients $L^i_{\ j}$ of Weingarten map

For a regular parametrisation $\underline{x}$ of a surface M, the vectors (x_1, x_2) from a basis of the tangent plane T_p. We can therefore exploit the uniqueness of the decomposition with respect to this basis, to define the coefficients of W_p as follows.

DEFINITION 9.4.1. The coefficients $L^i_{\ j}$ of the Weingarten map are defined by setting $W(x_j) = L^i_{\ j} x_i = L^1_{\ j}\, x_1 + L^2_{\ j}\, x_2$.

We summarize the the examples developed in the previous sections.

THEOREM 9.4.2. *For the plane, sphere, and cylinder one has:*

(1) *For the plane, we have* $(L^i_{\ j}) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$.

(2) *For the sphere, we have* $(L^i_{\ j}) = \begin{pmatrix} \frac{1}{r} & 0 \\ 0 & \frac{1}{r} \end{pmatrix} = \frac{1}{r}(\delta^i_{\ j})$.

(3) *For the cylinder, we have* $L^1_{\ 1} = 1$. *The remaining coefficients vanish, so that* $(L^i_{\ j}) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$.

9.5. Gaussian curvature

We consider the following data as in the previous sections:

(1) $M \subseteq \mathbb{R}^3$ is a surface;
(2) $\underline{x}(u^1, u^2)$ is a regular parametrisation of a neighborhood in M;
(3) $p \in M$ is a point of the surface;
(4) T_p is the tangent plane to M at p;
(5) x_1, x_2 are tangent vectors forming a basis for T_p;
(6) $W_p \colon T_p \to T_p$ is the Weingarten map at p;
(7) $L^i{}_j$ are the coefficients of W_p defined by $W_p(x_j) = L^i{}_j\, x_i$;
(8) the antisymmetrisation notation [ ] on indices was defined in Section 1.6.

Definition 9.5.1. The *Gaussian curvature* function on M, denoted $K = K(u^1, u^2)$, is the determinant of the Weingarten map at $p = \underline{x}(u^1, u^2)$:

$$K(u^1, u^2) = \det(W_p).$$

Corollary 9.5.2. *We have the formula*[1]

$$K = \det(L^i{}_j) = L^1{}_1 L^2{}_2 - L^1{}_2 L^2{}_1 = 2L^1{}_{[1} L^2{}_{2]}.$$

Corollary 9.5.3. *For both the cylinder and the plane we have the identity $K = 0$.*[2] *The Gaussian curvature of a sphere of radius r is $K = det \begin{pmatrix} \frac{1}{r} & 0 \\ 0 & \frac{1}{r} \end{pmatrix} = \dfrac{1}{r^2}$.*

Remark 9.5.4 (Sign of Gaussian curvature). Of particular geometric significance is the sign of the Gaussian curvature. The geometric meaning of negative Gaussian curvature is a saddle point. The geometric meaning of positive Gaussian curvature is a point of convexity, such as local minimum or local maximum of the graph of a function of two variables.

---

[1] An additional formula for K can be obtained via the signed curvatures $\tilde{k}_i$ of oriented curves, defined in Section 12.3 and analyzed in Section 12.7. One can then assert that Gaussian curvature K of a surface is the product $K = \tilde{k}_1 \cdot \tilde{k}_2$ of signed curvatures $\tilde{k}_1$ and $\tilde{k}_2$ of the pair of curves $M \cap P_1$ and $M \cap P_2$, where the planes P_i are defined by $P_i = \mathrm{Span}(n, v_i)$ and v_1, v_2 are orthogonal eigenvectors of the Weingarten map; see Theorem 12.10.4. We will mostly work with Definition 9.5.1 of Gaussian curvature; we mention the additional definition in terms of signed curvatures of curves as it was the original definition by Gauss.

[2] *cf.* Table 12.12.1.

9.6. Second fundamental form

In Section 8.8, we extended the normal vector field $n(u^1, u^2)$ along the surface $M \subseteq \mathbb{R}^3$ near a point $p \in M$, to a smooth vector field

$$N(x, y, z)$$

defined in an open neighborhood of p viewed as a point of $\mathbb{R}^3$, so that

$$N(\underline{x}(u^1, u^2)) = n(u^1, u^2). \tag{9.6.1}$$

DEFINITION 9.6.1. The second fundamental form $\mathbf{II}_p$ is the bilinear form on the tangent plane T_p defined for $u, v \in T_p$ by

$$\mathbf{II}_p(u, v) = -\langle \nabla_u N, v \rangle. \tag{9.6.2}$$

REMARK 9.6.2 (Surface geodesics as space curves). What does the second fundamental form measure? In the first approximation, one can say that the second fundamental form measures

the curvature of space curves defined by geodesics on M viewed as curves of $\mathbb{R}^3$ (cf. 4.2.2)

as illustrated by Theorem 9.7.1 below.

REMARK 9.6.3. The minus sign in formula (9.6.2) is just a convention, to make some later formulas, such as (9.6.3), to work out better. It is important to note the following:

the sign in (9.6.2) does not affect the sign of the Gaussian curvature!

DEFINITION 9.6.4. The coefficients L_{ij} of the second fundamental form are defined to be

$$L_{ij} = \mathbf{II}_p(x_i, x_j) = -\left\langle \frac{\partial n}{\partial u^i}, x_j \right\rangle.$$

LEMMA 9.6.5. *The coefficients L_{ij} of the second fundamental form are symmetric in i and j, more precisely*

$$L_{ij} = +\langle x_{ij}, n \rangle. \tag{9.6.3}$$

Thus we have $L_{12} = L_{21}$.

PROOF. By definition of the normal vector, we have $\langle n, x_i \rangle = 0$ everywhere. Hence $\frac{\partial}{\partial u^j} \langle n, x_i \rangle = 0$, *i.e.*

$$\left\langle \frac{\partial}{\partial u^j} n, x_i \right\rangle + \langle n, x_{ij} \rangle = 0.$$

Now by (9.6.1),

$$\langle n, x_{ij} \rangle = -\langle \nabla_{x_j} N, x_i \rangle = +\mathbf{II}_p(x_j, x_i)$$

and the proof is concluded by the equality of mixed partials (Clairaut–Schwarz theorem). $\square$

COROLLARY 9.6.6. *The second fundamental form enables us to identify the normal component of the second partial derivatives of the parametrisation* $\underline{x}(u^1, u^2)$. *In formulas,*

$$x_{ij} = \Gamma_{ij}^k x_k + L_{ij} n.$$

PROOF. We write $x_{ij} = \Gamma_{ij}^k x_k + cn$. Now form the inner product with n to obtain $L_{ij} = \langle x_{ij}, n \rangle = 0 + c\langle n, n \rangle = c$. $\square$

9.7. Geodesics and second fundamental form

The second fundamental form measures the curvature of geodesics on M viewed as space curves, in the following sense.

THEOREM 9.7.1. *Let* $\beta(s)$ *be a unit speed geodesic on a surface* $M \subseteq \mathbb{R}^3$, *so that* $\beta'(s) \in T_pM$ *at a point* $p = \beta(s)$. *Then the absolute value of the second fundamental form applied to the pair* (β', β') *is the curvature of* β *at* p:

$$|\mathbf{II}_p(\beta', \beta')| = k_\beta(s),$$

where k_β *is the curvature of the curve* β *viewed as a curve in* $\mathbb{R}^3$.

PROOF. We apply formula (7.2.3) for the curvature of a curve. Since $|n| = 1$ and β is a geodesic, we have

$$k_\beta \stackrel{\text{def}}{=} |\beta''| = \left| L_{ij} \alpha^{i'} \alpha^{j'} \right|. \tag{9.7.1}$$

On the other hand, recall that $\beta = \underline{x} \circ \alpha$. We have

$$\mathbf{II}_p(\beta', \beta') = \mathbf{II}_p(x_i \alpha^{i'}, x_j \alpha^{j'}) = \alpha^{i'} \alpha^{j'} \mathbf{II}_p(x_i, x_j) = \alpha^{i'} \alpha^{j'} L_{ij}.$$

This expression coincides with (9.7.1) up to sign, completing the proof. $\square$

9.8. Lowering and raising indices

PROPOSITION 9.8.1 (Lowering an index). *We have the following relation between the coefficients of the second fundamental form and the Weingarten map:*

$$L_{ij} = -L^k{}_j g_{ki}.$$

PROOF. By definition,

$$L_{ij} = \langle x_{ij}, n \rangle = -\left\langle \frac{\partial}{\partial u^j} n, x_i \right\rangle = -\langle L^k{}_j x_k, x_i \rangle = -L^k{}_j g_{ki}$$

proving the lemma. $\square$

COROLLARY 9.8.2 (Raising an index). *We have the following relation between the coefficients of the Weingarten map and the coefficients of the second fundamental form:* $L^i{}_j = -g^{ik} L_{kj}$.

PROOF. We start with the relation $L_{ij} = -L^k{}_j g_{ki}$ of Proposition 9.8.1. We now multiply the relation on both sides by $g^{i\ell}$, and sum over i, obtaining

$$\begin{aligned} L_{ij} g^{i\ell} &= -L^k{}_j g_{ki} g^{i\ell} \\ &= -L^k{}_j \delta_k^\ell \\ &= -L^\ell{}_j \end{aligned}$$

as required. □

9.9. Three formulas for Gaussian curvature

THEOREM 9.9.1. *We have the following three equivalent formulas for the Gaussian curvature:*

(a) $K = \det(L^i{}_j) = 2L^1{}_{[1} L^2{}_{2]}$;
(b) $K = \frac{\det(L_{ij})}{\det(g_{ij})}$;[3]
(c) $K = -\frac{2}{g_{11}} L_{1[1} L^2{}_{2]}$.

PROOF. Here formula (a) is our definition of K. To prove formula (b), we use the formula

$$L_{ij} = -L^k{}_j g_{ki} \tag{9.9.1}$$

of Proposition 9.8.1. By the multiplicativity of determinant with respect to matrix multiplication,

$$\det(L^i{}_j) = (-1)^2 \frac{\det(L_{ij})}{\det(g_{ij})}.$$

Let us now prove formula (c). The proof is a calculation. Note that by definition of the antisymmetrisation notation, $2L_{1[1} L^2{}_{2]} = L_{11} L^2{}_2 - L_{12} L^2{}_1$. Applying (9.9.1), we obtain

$$\begin{aligned} -\frac{2}{g_{11}} \left(L_{1[1} L^2{}_{2]} \right) &= \frac{1}{g_{11}} \left(\left(L^1{}_1 g_{11} + L^2{}_1 g_{21} \right) L^2{}_2 - \left(L^1{}_2 g_{11} + L^2{}_2 g_{21} \right) L^2{}_1 \right) \\ &= \frac{1}{g_{11}} \left(g_{11} L^1{}_1 L^2{}_2 + g_{21} L^2{}_1 L^2{}_2 - g_{11} L^1{}_2 L^2{}_1 - g_{21} L^2{}_2 L^2{}_1 \right) \\ &= \det(L^i{}_j) = K \end{aligned}$$

[3]It is important to note that the sign of K is not affected by the sign convention we adopted in (9.6.2) when we defined the second fundamental form. The formula $K = \det(L_{ij})/\det(g_{ij})$ would be correct with either sign convention there.

as required.[4] □

9.10. Principal curvatures and geodesics

We start with the following data.

(1) a point $p \in M$ of a surface M.
(2) $W_p \colon T_p \to T_p$ is an endomorphism of the tangent plane T_pM.
(3) The principal curvatures at p were defined in Sections 8.12 and 8.14.

Let us recall the definition of the principal curvatures.

DEFINITION 9.10.1. The *principal curvatures of M at the point p*, denoted k_1 and k_2, are the eigenvalues of the Weingarten map W_p.

PROPOSITION 9.10.2. *Let M be the hyperbolic paraboloid given by the graph in $\mathbb{R}^3$ of the function $f(x, y) = ax^2 - by^2$, $a, b > 0$.*[5] *Then the principal curvatures at the origin are $k_1 = 2a$ and $k_2 = -2b$.*[6]

PROOF. The origin $p = (0, 0)$ is a critical point of f. Therefore the matrix of the Weingarten map W_p at p is the Hessian matrix H_f of f, namely $H_f = \begin{pmatrix} 2a & 0 \\ 0 & -2b \end{pmatrix}$. □

See further in Section 12.11 on the connection between the Weingarten map and the Hessian.

REMARK 9.10.3. The curvatures k_1 and k_2 are necessarily real, by the selfadjointness of W (Theorem 8.14.1) together with Corollary 2.3.2.

COROLLARY 9.10.4. *The Gaussian curvature $K(u^1, u^2)$ of M at a point $p = \underline{x}(u^1, u^2)$ equals the product of the principal curvatures at p.*

PROOF. The determinant of a 2 by 2 matrix equals the product of its eigenvalues: $K = \det(L^i{}_j) = k_1 k_2$, where $(L^i{}_j)$ is the matrix representing the endomorphism W_p. □

Recall that $\mathbf{II}_p$ denotes the second fundamental form at $p \in M$. We proved in Theorem 9.7.1 that

$$|\mathbf{II}_p(\beta', \beta')| = k_\beta, \tag{9.10.1}$$

---

[4] In the proof above, we could calculate using either the formula $L_{ij} = -L^k{}_j g_{ki}$, or the formula $L_{ij} = -L^k{}_i g_{kj}$. Only the former one leads to the appropriate cancellations as above.

[5] See Section 3.6.

[6] The order of the two eigenvalues is immaterial.

where k_β is the curvature of the unit speed geodesic $\beta(s)$ with velocity vector β'. We now apply this result to the eigendirections.

THEOREM 9.10.5. *Let k be a principal curvature at the point $p \in M$. Let $\beta(s)$ be a geodesic on M satisfying $\beta(0) = p$, such that $\beta'(0)$ is an eigenvector belonging to k. Then the curvature of β as a space curve is the absolute value of k :*

$$k_\beta(0) = |k|.$$

PROOF OF THEOREM 9.10.5. Since β' is an eigenvector of W_p, we obtain from (9.10.1) that

$$\begin{aligned} k_\beta(0) &= |\mathbf{II}_p(\beta', \beta')| \\ &= \left| \left\langle W_p\,(\beta'), \beta' \right\rangle \right| \\ &= |\langle k\beta', \beta' \rangle| \\ &= |k\langle \beta', \beta' \rangle| \\ &= |k| \end{aligned}$$

proving the theorem.[7] $\square$

See further in Theorem 12.10.4.

9.11. Mean curvature, minimal surfaces

We start with the following data.

(1) $p \in M$ a point on a surface M;
(2) T_p, the tangent plane at p;
(3) $W_p : T_p \to T_p$ the Weingarten map, an endomorphism of T_p.

DEFINITION 9.11.1. The *mean curvature* $H = H(u^1, u^2)$ of a surface $M \subseteq \mathbb{R}^3$ at a point $p = \underline{x}(u^1, u^2)$ is half the trace of the Weingarten map:

$$H = \tfrac{1}{2}\text{trace}(W_p) = \tfrac{1}{2}(k_1 + k_2) = \tfrac{1}{2}L^i{}_i.$$

DEFINITION 9.11.2. A surface $M \subseteq \mathbb{R}^3$ is called *minimal* if $H = 0$ at every point, *i.e.* $k_1 + k_2 = 0$.

[7]The absolute value of the Gaussian curvature at a point p of a regular surface in $\mathbb{R}^3$ is the product of curvatures of two perpendicular geodesics passing through p, whose tangent vectors are eigenvectors of the Weingarten map at the point. The absolute value can be removed. Choose a unit normal n to the surface M at $p \in M$. Consider geodesics $\beta_i(s)$ such that $\beta_1'(0)$ and $\beta_2'(0)$ are the principal directions at p. We can define the signed curvature $\tilde{k}_i$ of geodesics $\beta_i(s)$ by setting $\tilde{k}_i = \langle \beta_i, n \rangle$. Then $K = k_1 k_2$ (without absolute value bars).

Remark 9.11.3 (Meaning of sign). The *sign* of the mean curvature has no geometric meaning and depends on the choice of normal vector n (from among the pair $n, -n$) used in the definition of W_p and $\mathbf{II}_p$. This is in contrast with Gaussian curvature of M where the sign does have geometric meaning.

See Table 12.12.1 on plane and cylinder for comparison of the geometric meaning of mean curvature and Gaussian curvature.

Remark 9.11.4 (Area minimisation; soap films). Geometrically, a minimal surface is represented locally by a soap film.[8] A soap film seeks to minimize area locally. In this sense minimal surfaces are similar to geodesics, which minimize length locally.

9.11.1. Scherk surface. This material is optional. The Scherk surface $M_S \subseteq \mathbb{R}^3$ is a well-known minimal surface which is periodic in the sense explained below.

Definition 9.11.5. The Scherk surface in $M_S \subseteq \mathbb{R}^3$ is the graph of $z = \ln \frac{\cos y}{\cos x}$.

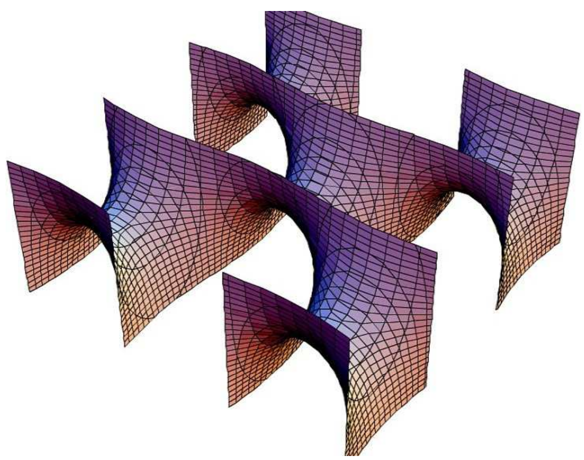

Figure 9.11.1. Scherk surface

Thus M_S is parametrized by $\left(x, y, \ln\left(\frac{\cos y}{\cos x}\right)\right)$. Since cosine is periodic, so is the parametrisation in both variables.

Lemma 9.11.6. *The matrix of the first fundamental form of M_S is diagonal.*

[8] krum sabon, as opposed to bu'at sabon. Dip (tvol) a wire (tayil) into soapy water (mei sabon).

PROOF. The parametrisation $\underline{x}(x, y) = (x, y, \ln\cos y - \ln\cos x)$ of M_S clearly satisfies $\underline{x}_{12} = 0$. Therefore $L_{12} = \langle \underline{x}_{12}, n\rangle = 0$. Thus the matrix (L_{ij}) is diagonal. □

THEOREM 9.11.7 (Scherk surface). *The Scherk surface is a minimal surface.*

PROOF. The proof is in 3 steps.

Step 1. By Corollary 9.8.2 (raising an index), we have $L^1_1 = -L_{11}g^{11}$ since the first fundamental form is diagonal by Lemma 9.11.6. Similarly, $L^2_2 = -L_{22}g^{22}$. Thus, the mean curvature H satisfies the formula $2H = \operatorname{trace} W_p = -(L_{11}g^{11} + L_{22}g^{22})$. Therefore the condition $H = 0$ is equivalent to

$$L_{11}g^{11} + L_{22}g^{22} = 0. \tag{9.11.1}$$

Let $g = \det(g_{ij})$.[9] By definition of the inverse matrix, $g^{11} = \frac{g_{22}}{g}$ and $g^{22} = \frac{g_{11}}{g}$. Thus the condition (9.11.1) becomes $\frac{L_{11}g_{22}+L_{22}g_{11}}{g} = 0$ where $g \neq 0$, or equivalently

$$L_{11}g_{22} + L_{22}g_{11} = 0.$$

Thus we have the following reformulation of minimality for M_S:

$$M_S \text{ is minimal } \longleftrightarrow \frac{L_{11}}{g_{11}} + \frac{L_{22}}{g_{22}} = 0. \tag{9.11.2}$$

Step 2. For the partial derivatives of $\underline{x}(x, y) = (x, y, \ln\cos y - \ln\cos x)$, we have

$$\begin{cases} x_1 = (1, 0, -\tan x)^t, \\ x_2 = (0, 1, \tan y)^t. \end{cases} \tag{9.11.3}$$

Hence

$$g_{11} = 1 + \tan^2 x = \sec^2 x, \quad g_{22} = 1 + \tan^2 y = \sec^2 y. \tag{9.11.4}$$

The normal vector is the normalisation of the vector product $x_1 \times x_2$. The vector product is $(-\tan x, \tan y, 1)^t$. Let $C = \sqrt{1 + \tan^2 x + \tan^2 y}$, so that

$$n = \tfrac{1}{C}(\tan x, -\tan y, 1)^t.$$

Step 3. We have $x_{11} = (0, 0, -\sec^2(x))^t$ and $x_{22} = (0, 0, \sec^2 y)^t$. Therefore

$$\begin{cases} L_{11} = \langle n, \underline{x}_{11}\rangle = -\frac{\sec^2 x}{C}, \\ L_{22} = \langle n, \underline{x}_{22}\rangle = \frac{\sec^2 y}{C}. \end{cases} \tag{9.11.5}$$

Thus by (9.11.4), $\frac{L_{11}}{g_{11}} = -\frac{1}{C}$ and $\frac{L_{22}}{g_{22}} = \frac{1}{C}$, and therefore $\frac{L_{11}}{g_{11}} + \frac{L_{22}}{g_{22}} = 0$. Hence by (9.11.2), M is minimal, proving the minimality of the Scherk surface. □

See Figure 9.11.1 for Scherk surface.

[9] Note that $g_{12} \neq 0$ in this case.

REMARK 9.11.8. The standard parametrisation $\left(x, y, \ln \frac{\cos y}{\cos x}\right)$ of the Scherk surface provides an example of a parametrisation where the second fundamental form is given by a diagonal matrix, but the first fundamental form and the Weingarten map are given by nondiagonal matrices.

See Figure 10.6.2 for another example of a minimal surface, called the helicoid.

CHAPTER 10

Hyperbolic plane; minimal surf; Theorema egregium

10.1. Metrics conformal to the flat metric and their Γs

A particularly important class of metrics are those conformal to the flat metric, in the following sense. We will use the symbol

$$\lambda = \lambda(u^1, u^2)$$

for the conformal factor of the metric, as below.

DEFINITION 10.1.1. A metric (g_{ij}) is *conformal* to the standard flat metric if there is a function $\lambda(u^1, u^2) > 0$ such that $g_{ij}(u^1, u^2) = \lambda\,\delta_{ij}$ for all $i, j = 1, 2$.

In other words, the matrix of the first fundamental form is a scalar matrix at each point. We will use the notation $\lambda_i = \frac{\partial \lambda}{\partial u^i}$.

LEMMA 10.1.2. *Let $g_{ij} = \lambda(u^1, u^2)\,\delta_{ij}$ be a metric conformal to the standard flat metric. Then*

$$\begin{cases} \Gamma_{11}^1 = \frac{\lambda_1}{2\lambda}, \\ \Gamma_{22}^1 = \frac{-\lambda_1}{2\lambda}, \\ \Gamma_{12}^1 = \frac{\lambda_2}{2\lambda}. \end{cases}$$

The values of the coefficients are listed in Table 10.1.1.

Γ_{ij}^1	$j = 1$	$j = 2$
$i = 1$	$\frac{\lambda_1}{2\lambda}$	$\frac{\lambda_2}{2\lambda}$
$i = 2$	$\frac{\lambda_2}{2\lambda}$	$\frac{-\lambda_1}{2\lambda}$

Γ_{ij}^2	$j = 1$	$j = 2$
$i = 1$	$-\frac{\lambda_2}{2\lambda}$	$\frac{\lambda_1}{2\lambda}$
$i = 2$	$\frac{\lambda_1}{2\lambda}$	$\frac{\lambda_2}{2\lambda}$

TABLE 10.1.1. Symbols Γ_{ij}^k of a conformal metric $\lambda\,\delta_{ij}$

PROOF. We will use the general formula proved in Section 6.6:

$$\Gamma_{ij}^{k} = \frac{1}{2}(g_{i\ell,j} - g_{ij,\ell} + g_{j\ell,i})g^{\ell k}. \tag{10.1.1}$$

For diagonal metrics, formula (10.1.1) simplifies to

$$\Gamma_{ij}^{k} = \frac{1}{2}(g_{i\underline{k},j} - g_{ij,\underline{k}} + g_{j\underline{k},i})g^{\underline{k}\,\underline{k}}. \tag{10.1.2}$$

We have underlined the index k in (10.1.2) to emphasize that no summation is taking place even though k appears as both a subscript and a superscript. Equivalently, we can write

$$\Gamma_{ij}^{k} = \frac{1}{2g_{kk}}(g_{ik,j} - g_{ij,k} + g_{jk,i})$$

where underlining is no longer necessary since the index k appears only as a subscript in the formula on the right-hand side. If $i = j$ then the formula simplifies to

$$\Gamma_{ii}^{k} = \frac{1}{2g_{kk}}(g_{ik,i} - g_{ii,k} + g_{ik,i}) = \frac{1}{2g_{kk}}(2g_{ik,i} - g_{ii,k}).$$

By hypothesis, we have $g_{11} = g_{22} = \lambda(u^1, u^2)$ while $g_{12} = 0$ and the lemma follows by examining the cases. If $i = j = 1$ then

$$\Gamma_{11}^{1} = \frac{g_{11,1}}{2\lambda} = \frac{\lambda_1}{2\lambda}.$$

If $i = j = 2$ then

$$\Gamma_{22}^{1} = \frac{2 \cdot 0 - g_{22,1}}{2\lambda} = -\frac{\lambda_1}{2\lambda}.$$

Similar calculations yield the formulas for the coefficients Γ_{ij}^{2} which are all listed in the Table. □

An important example is the calculation of the geodesic equation of the hyperbolic metric of Section 10.3.

10.2. The hyperbolic plane

Let $x = u^1$ and $y = u^2$.

DEFINITION 10.2.1. The upper half-plane is $\mathcal{H}^2 = \{(x, y) \colon y > 0\}$.

THEOREM 10.2.2. *The metric with coefficients*

$$g_{ij} = \frac{1}{y^2}\delta_{ij} \tag{10.2.1}$$

in the upper half-plane $\mathcal{H}^2$ has constant Gaussian curvature $K = -1$.

PROOF. In Theorem 11.2.2, we will prove that $K = -\Delta_{LB} \ln f$ where $\Delta_{LB} = \frac{1}{f^2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial x^2}\right)$. Hence

$$K = -\Delta_{LB} \ln f = \Delta_{LB} \ln y = y^2 \left(-\frac{1}{y^2}\right) = -1,$$

as required. □

DEFINITION 10.2.3. The metric (10.2.1) is called the *hyperbolic metric* of the upper half-plane. The upper half-plane equipped with the hyperbolic metric is called the *hyperbolic plane*.

This example is also discussed in Section 16.20.

10.3. Geodesics in the hyperbolic plane

DEFINITION 10.3.1 (Isothermal coordinates). Coordinates (u^1, u^2) with respect to which the metric is expressed by $g_{ij} = \lambda(u^1, u^2)\delta_{ij}$ are called *isothermal coordinates*.

See Section 10.4 for additional details.

LEMMA 10.3.2. *For the hyperbolic metric we have*

$$\begin{cases} \Gamma^1_{11} = \Gamma^1_{22} = 0 \\ \Gamma^1_{12} = -\frac{1}{y}. \end{cases} \tag{10.3.1}$$

PROOF. In Section 10.1 we computed the Gamma symbols of a metric relative to isothermal coordinates. We have $\Gamma^1_{11} = \frac{\lambda_1}{2\lambda}$, $\Gamma^1_{22} = \frac{-\lambda_1}{2\lambda}$, and $\Gamma^1_{12} = \frac{\lambda_2}{2\lambda}$. Since $\lambda(x, y) = y^{-2}$, we obtain $\Gamma^1_{11} = \Gamma^1_{22} = 0$ and $\Gamma^1_{12} = \frac{\lambda_2}{2\lambda} = \frac{-2y^{-3}}{2y^{-2}} = -\frac{y^2}{y^3} = -\frac{1}{y}$. □

THEOREM 10.3.3 (Geodesics in the hyperbolic plane). *The differential equations of a geodesic* $(x(s), y(s))$ *in the hyperbolic plane are*

$$\begin{cases} y\,x'' - 2x'y' = 0 & \text{for } k = 1 \\ y\,y'' + (x')^2 - (y')^2 = 0 & \text{for } k = 2. \end{cases}$$

PROOF. The differential equation of a geodesic is

$$\alpha^{k''} + \Gamma^k_{ij}\alpha^{i'}\alpha^{j'} = 0.$$

For $k = 1$, by Lemma 10.3.2 this becomes

$$\alpha^{1''} + 2\Gamma^1_{12}\alpha^{1'}\alpha^{2'} = 0.$$

Since $\alpha^1(s) = x(s)$ and $\alpha^2(s) = y(s)$, we obtain $x'' + 2\Gamma^1_{12}x'y' = 0$ which using formula (10.3.1) of Lemma 10.3.2 becomes

$$x'' - \frac{2}{y}x'y' = 0.$$

Now let $k = 2$. We have

$$\begin{cases} \Gamma^2_{11} = -\frac{\lambda_2}{2\lambda} = \frac{1}{y} \\ \Gamma^2_{22} = -\frac{1}{y} \\ \Gamma^2_{12} = \frac{\lambda_1}{2\lambda} = 0. \end{cases}$$

Therefore the geodesic equation for $k = 2$ is $y'' + \frac{1}{y}x'x' - \frac{1}{y}y'y' = 0$. □

EXAMPLE 10.3.4 (Vertical ray). The exponential function $y = e^s$ (while x is constant as a function of s) satisfies both equations of Theorem 10.3.3. Indeed, for a vertical line $y = y(t)$, the first equation is trivially satisfies. The second equation gives $yy'' - (y')^2 = 0$ or

$$\left(\frac{y'}{y}\right)' = 0.$$

Therefore $\frac{y'}{y} = C$, or $\frac{dy}{ds} = Cy$. Separating the variables we obtain $\frac{dy}{y} = Cds$. Integrating, we obtain $\ln y = Cs$, or $y = e^{Cs}$. Therefore the formula

$$y(s) = e^{Cs}$$

provides a constant speed parametrisation of a hyperbolic geodesic which traces out a vertical half-line in the upper half-plane. Since the exponential function is positive, the geodesic never reaches the "boundary" x-axis.[1]

EXAMPLE 10.3.5. The pseudosphere of Section 5.9 is an example of a surface of constant Gaussian curvature -1 embedded in Euclidean space; see Section 13.3. It is therefore locally isometric to the hyperbolic plane.

COROLLARY 10.3.6 (An identity for the Gamma symbols). *A metric in isothermal coordinates* (u^1, u^2) *satisfies the following identities:*

$$\begin{cases} \Gamma^1_{11} + \Gamma^1_{22} = 0 \\ \Gamma^2_{11} + \Gamma^2_{22} = 0. \end{cases}$$

PROOF. From Table 10.1.1 we have $\Gamma^1_{11} + \Gamma^1_{22} = \frac{\lambda_1}{2\lambda} - \frac{\lambda_1}{2\lambda} = 0$. Similarly, $\Gamma^2_{11} + \Gamma^2_{22} = -\frac{\lambda_2}{2\lambda} + \frac{\lambda_2}{2\lambda} = 0$. □

This corollary will be used in Section 10.4 to relate the Laplacian to the mean curvature.

The Gaussian curvature takes a particularly simple form with respect to isothermal coordinates; see Section 11.2.

[1]The same is true of all geodesics in the hyperbolic plane. By the Hopf–Rinow theorem, this is equivalent to the *completeness* property of the hyperbolic plane.

10.4. Minimal surfaces in isothermal coordinates

In this section we study the partial differential equation (PDE) of a minimal surface $M \subseteq \mathbb{R}^3$. The following definition recalls Definition 10.3.1.

DEFINITION 10.4.1. A parametrisation $\underline{x}(u^1, u^2)$ is called *isothermal* if the following two equivalent conditions are satisfied:

(1) there is a function $f = f(u^1, u^2) > 0$ such that $g_{ij} = f^2\delta_{ij}$,
(2) $\langle x_1, x_1\rangle = \langle x_2, x_2\rangle$ and $\langle x_1, x_2\rangle = 0$.

DEFINITION 10.4.2. The function f^2 is called the *conformal factor* of the metric.

Sometimes the function f itself is called the conformal factor. The usage will be clear from context.

PROPOSITION 10.4.3. *A surface $M \subseteq \mathbb{R}^3$ with parametrisation $\underline{x}(u^1, u^2)$ in isothermal coordinates satisfies the PDE*

$$\Delta\underline{x} = -2f^2H\underline{n},$$

where $H = H(u^1, u^2)$ is the mean curvature and $n = n(u^1, u^2)$ is the normal vector to $M \subseteq \mathbb{R}^3$.

PROOF. We use the formula $x_{ij} = \Gamma^k_{ij}x_k + L_{ij}n$ to write

$$\begin{aligned}\Delta\,\underline{x} &= x_{11} + x_{22}\\ &= \Gamma^1_{11}x_1 + \Gamma^2_{11}x_2 + L_{11}n + \Gamma^1_{22}x_1 + \Gamma^2_{22}x_2 + L_{22}n\\ &= \left(\Gamma^1_{11} + \Gamma^1_{22}\right)x_1 + \left(\Gamma^2_{11} + \Gamma^2_{22}\right)x_2 + (L_{11} + L_{22})\,n.\end{aligned}$$

By Corollary 10.3.6, in isothermal coordinates we have $\Gamma^1_{11} + \Gamma^1_{22} = 0$ and $\Gamma^2_{11} + \Gamma^2_{22} = 0$. Hence $\Delta\,\underline{x} = (L_{11} + L_{22})\,n$. Now recall that with respect to isothermal coordinates, we have $L_{ii} = -f^2L^{\underline{i}}_{\underline{i}}$ for each i. Therefore $\Delta\,\underline{x} = (L_{11} + L_{22})n = -(L^1_1 + L^2_2)f^2n = -2Hf^2n$, as required. [2] $\square$

[2] An alternative proof following Do Carmo [**Ca76**]. This material is optional. Note that $L_{ij} = -L^m{}_j g_{mi} = -L^m{}_j f^2\delta_{mi} = -f^2L^i{}_j$. Thus $L^i{}_j = -\frac{L_{ij}}{f^2}$ so that the mean curvature H satisfies

$$H = \frac{1}{2}L^i{}_i = -\frac{L_{11} + L_{22}}{2f^2}. \tag{10.4.1}$$

Differentiating $\langle x_1, x_2\rangle = 0$ we obtain $\frac{\partial}{\partial u^2}\langle x_1, x_2\rangle = 0$. Therefore

$$\langle x_{12}, x_2\rangle + \langle x_1, x_{22}\rangle = 0. \tag{10.4.2}$$

So $-\langle x_{12}, x_2\rangle = \langle x_1, x_{22}\rangle$. By Definition 10.4.1 of isothermal coordinates, we have

$$\langle x_1, x_1\rangle - \langle x_2, x_2\rangle = 0. \tag{10.4.3}$$

10.5. Minimality and harmonic functions

In this section we study the relation between minimal surfaces and harmonic functions. The latter are familiar from complex function theory. The Laplacian was defined in Section 4.6, and already used in Section 10.4.

DEFINITION 10.5.1. We say that F is *harmonic* if $\Delta F = 0$.

Harmonic functions are important in the study of heat flow, or heat transfer.[3]

THEOREM 10.5.2. *Let $M \subseteq \mathbb{R}^3$ be a surface with parametrisation $\underline{x}(u^1, u^2) = (x(u^1, u^2), y(u^1, u^2), z(u^1, u^2))$ Assume that the coordinates (u^1, u^2) are isothermal. Then M is* minimal *if and only if the coordinate functions x, y, z are* harmonic.

PROOF. From Proposition 10.4.3 we have

$$|\Delta(\underline{x})| = \sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2} = 2f^2|H|.$$

Therefore the Laplacian of $\underline{x}$ vanishes if and only if $\Delta x = \Delta y = \Delta z = 0$, which occurs if and only if $H = 0$. $\square$

10.6. Catenoid is minimal

THEOREM 10.6.1 (Catenoid is a minimal surface). *Let $a > 0$. The* catenoid *parametrized by*

$$\underline{x}(\theta, \phi) = (a \cosh \phi \cos \theta, a \cosh \phi \sin \theta, a\phi),$$

is a minimal surface.

PROOF. The generating curve is the curve $r(\phi) = a \cosh \phi$ and $z(\phi) = a\phi$.[4] Then according to the general formula for a surface of revolution, $g_{11} = r^2 = a^2 \cosh^2 \phi$. Also,

$$g_{22} = (\tfrac{dr}{d\phi})^2 + (\tfrac{dz}{d\phi})^2 = (a \sinh \phi)^2 + a^2 = a^2 \cosh^2 \phi = g_{11},$$

Differentiating (10.4.3) and applying (10.4.2), we obtain $0 = \frac{\partial}{\partial u^1}\langle x_1, x_1\rangle - \frac{\partial}{\partial u^1}\langle x_2, x_2\rangle = 2\langle x_{11}, x_1\rangle - 2\langle x_{21}, x_2\rangle = 2\langle x_{11}, x_1\rangle + 2\langle x_{22}, x_1\rangle = 2\langle x_{11} + x_{22}, x_1\rangle$. Inspecting the u^2-derivatives, we similarly obtain $\langle x_{11} + x_{22}, x_2\rangle = 0$. Since x_1, x_2 and n form an orthogonal basis, the sum $x_{11} + x_{22}$ is proportional to n. Write $x_{11} + x_{22} = cn$. Applying (10.4.1), we obtain $c = \langle x_{11} + x_{22}, n\rangle = \langle x_{11}, n\rangle + \langle x_{22}, n\rangle = L_{11} + L_{22} = -2f^2 H$, as required.

[3] ma'avar chom.

[4] This is the catenary curve expressing the shape of suspension bridges.

and $g_{12} = 0$. We conclude that the coordinates (θ, ϕ) are isothermal. Finally,

$$\begin{aligned} x_{11} + x_{22} &= (-a\cosh\phi\cos\theta, -a\cosh\phi\sin\theta, 0)^t \\ &\quad + (a\cosh\phi\cos\theta, a\cosh\phi\sin\theta, 0)^t \\ &= (0,0,0). \end{aligned}$$

By Theorem 10.5.2, we have $H = 0$ and therefore the catenoid is a minimal surface. □

See Figure 10.6.1.

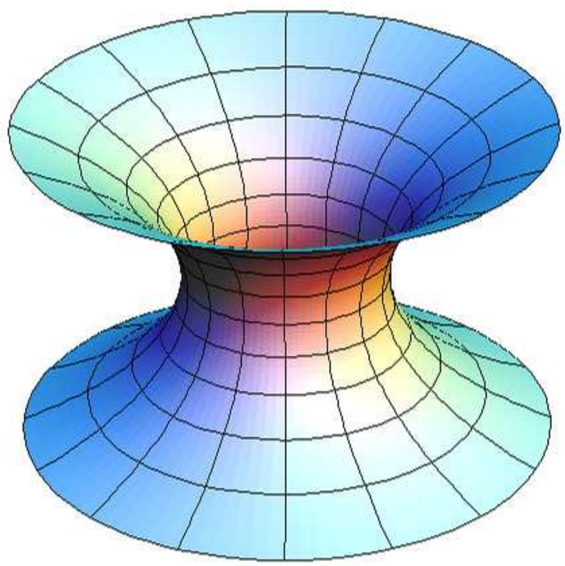

FIGURE 10.6.1. Catenoid: a minimal surface

REMARK 10.6.2. The catenoid is the only complete surface of revolution which is minimal [**We55**, p. 179], other than the plane. The helicoid (see Figure 10.6.2) and the catenoid are locally isometric.

10.7. Intro to theorema egregium

Understanding the intrinsic nature of Gaussian curvature, *i.e.* the *theorema egregium* of Gauss, clarifies the geometric classification of

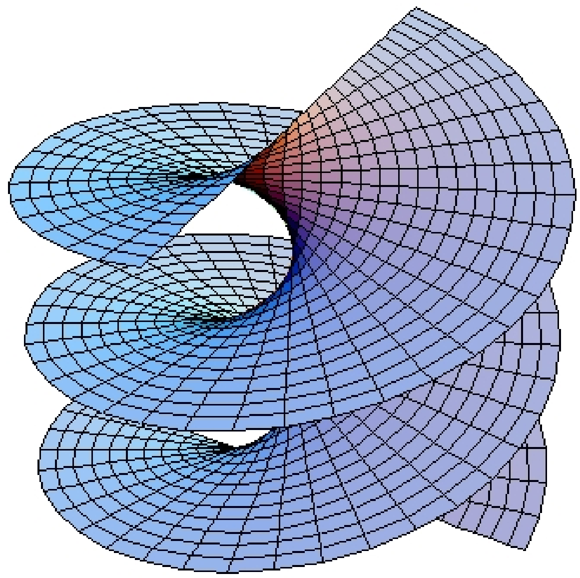

FIGURE 10.6.2. Helicoid: a minimal surface

surfaces.[5] A historical account and an analysis of Gauss's proof of the *theorema egregium* may be found in Dombrowski [**Do79**, p. 105].[6]

We will present a compact formula for Gaussian curvature in Section 10.10, and its proof, along the lines of the argument in Manfredo do Carmo's book [**Ca76**].[7] The appeal of an old-fashioned, computational, coordinate notation proof is that it obviates the need for higher order objects such as connections, tensors, exponental map, *etc*, and

[5]Our formula (10.10.2) for Gaussian curvature is similar to the traditional formula for the Riemann curvature tensor in terms of the Levi-Civita connection (note the antisymmetrisation in both formulas, and the corresponding two summands), without the burden of the connection formalism.

[6]The background, and the applications, have to do with practical surveying of the earth's surface. One of the uses Gauss makes of the *theorema egregium* is a comparison theorem for small triangles when one compares a triangle in the surface with the corresponding triangle of the same sidelengths in the plane. This is discussed in Dombrowski, pages 114–115. The earth is less curved toward the poles, with the result that the angular correction will be smaller for vertices closer to the poles: "Gauss gives these different correction values according to (35) for one of the largest terrestial triangles measured by him, namely with 'vertices' at Brocken, Hohehagen and Inselsberg" (Dombrowski p. 115).

[7]Other undergraduate differential geometry textbooks are discussed at `https://mathoverflow.net/questions/7834/undergraduate-differential-geometry-texts`

is, hopefully, directly accessible to a student not yet familiar with the subject (see note 14).

REMARK 10.7.1. We would like to distinguish two types of properties of a surface in Euclidean space: intrinsic and extrinsic.

Understanding the extrinsic/intrinsic dichotomy is equivalent to understanding the *theorema egregium* of Gauss. The *theorema egregium* is the key insight lying at the foundation of differential geometry as conceived by Bernhard Riemann in his essay [**Ri1854**] presented before the Royal Scientific Society of Göttingen in 1854 (see Section 10.7).

The *theorema egregium* asserts that an infinitesimal invariant of a surface in Euclidean space, called Gaussian curvature, is an "intrinsic" invariant of the surface M. In other words, Gaussian curvature of M is independent of its isometric embedding in Euclidean space. This theorem paves the way for an intrinsic definition of curvature in modern Riemannian geometry.

REMARK 10.7.2. We will prove that Gaussian curvature K is an *intrinsic* invariant of a surface in Euclidean space, in the following precise sense: K can be expressed in terms of the coefficients of the *first* fundamental form and their derivatives alone.

A priori the possibility of expressing K in such a fashion is not obvious, as the naive definition of K is in terms of the Weingarten map.

The intrinsic nature of Gaussian curvature paves the way for a transition from classical differential geometry, to a more abstract approach of modern differential geometry (see also Subsection 13.4). The distinction can be described roughly as follows. Classically, one studies surfaces in Euclidean space. Here the first fundamental form (g_{ij}) of the surface is the restriction of the Euclidean inner product. Meanwhile, abstractly, a surface comes equipped with a set of coefficients, which we deliberately denote by the same letters, (g_{ij}) in each coordinate patch, or equivalently, its element of length. One then proceeds to study its geometry without any reference to a Euclidean embedding, *cf.* (16.7.2).

10.8. Riemann's formula

Such an approach was pioneered in higher dimensions in Riemann's essay. The essay contains a single formula [**Ri1854**, p. 292]. See Spivak [**Sp79**, p. 149]), where the formula appears on page 159. This is the formula for the element of length of a surface of constant (Gaussian) curvature $K \equiv \alpha$:

$$\frac{1}{1+\frac{\alpha}{4}\sum x^2}\sqrt{\sum dx^2} \tag{10.8.1}$$

where today, of course, we would incorporate a summation index i as part of the notation, as in

$$ds = \frac{1}{1+\frac{\alpha}{4}\sum_i (x^i)^2}\sqrt{\sum_i (dx^i)^2} \tag{10.8.2}$$

cf. formulas (16.7.3), (16.20.1).

We will explain the notation dx^i in Section 13.9.

Remark 10.8.1. As noted in earlier sections, given a surface M in 3-space which is the graph of a function of two variables, at a *critical point* $p \in M$, the *Gaussian curvature* of the surface at the critical point p is the determinant of the Hessian of the function, i.e., the determinant of the two-by-two matrix of its second derivatives.

The implicit function theorem allows us to view any point of a regular surface, as such a critical point, after a suitable rotation in $\mathbb{R}^3$. We have thus given the simplest possible definition of Gaussian curvature at any point of a regular surface.

Remark 10.8.2. The appeal if this definition is that it allows one immediately to grasp the basic distinction between negative *versus* positive curvature, in terms of the dichotomy "saddle point *versus* cup/cap".

It will be more convenient to use an alternative definition in terms of the Weingarten map, which is readily shown to be equivalent to the definition in terms of the Hessian (see further in Section 12.11).

(1) (u^1, u^2) are coordinates in $\mathbb{R}^2$.
(2) $\underline{x} = \underline{x}(u^1, u^2)\colon \mathbb{R}^2 \to \mathbb{R}^3$ is a regular parametrized surface (i.e., the vector valued function $\underline{x}$ has Jacobian of rank 2 at every point).
(3) We have the partial derivatives $x_i = \frac{\partial}{\partial u^i}(\underline{x})$, where $i = 1, 2$. Thus, vectors x_1 and x_2 form a basis for the tangent plane at every point.
(4) We let $x_{ij} = \frac{\partial^2 x}{\partial u^i \partial u^j} \in \mathbb{R}^3$.
(5) $\langle\, ,\rangle$ is the Euclidean inner product.
(6) We let $n = n(u^1, u^2)$ be a unit normal to the surface at the point $x(u^1, u^2)$, so that $\langle n, x_i\rangle = 0$.
(7) The first fundamental form (g_{ij}) is given in coordinates by $g_{ij} = \langle x_i, x_j\rangle$.

(8) The second fundamental form (L_{ij}) is given in coordinates by $L_{ij} = \langle n, x_{ij} \rangle$.
(9) The symbols Γ^k_{ij} are uniquely defined by the decomposition

$$\begin{aligned} x_{ij} &= \Gamma^k_{ij} x_k + L_{ij} n \\ &= \Gamma^1_{ij} x_1 + \Gamma^2_{ij} x_2 + L_{ij} n, \end{aligned} \tag{10.8.3}$$

where the repeated (upper and lower) index k implies summation, in accordance with the Einstein summation convention.
(10) The Weingarten map $(L^i{}_j)$ is an endomorphism of the tangent plane $\mathbb{R}x_1 + \mathbb{R}x_2$. It is uniquely defined by the decomposition

$$\begin{aligned} n_j &= L^i{}_j x_i \\ &= L^1{}_j x_1 + L^2{}_j x_2, \end{aligned}$$

where $n_j = \frac{\partial}{\partial u^j}(n)$.
(11) We have the relation $L_{ij} = -L^k{}_i \, g_{jk}$.
(12) We use the notation $\Gamma^k_{ij,\ell}$ for the ℓ-th partial derivative of the symbol Γ^k_{ij}.
(13) We will denote by square brackets [ ] the antisymmetrisation over the pair of indices found in between the brackets, *e.g.* $a_{[ij]} = \frac{1}{2}(a_{ij} - a_{ji})$.
(14) We have $g_{[ij]} = 0$, $L_{[ij]} = 0$, $\Gamma^k_{[ij]} = 0$, and $x_{[ij]} = 0$.

10.9. An identity involving the Γs and the *L*s

Let $\underline{x}(u^1, u^2)$ be a regular parametrisation of a surface in $\mathbb{R}^3$.

LEMMA 10.9.1. *The third partial derivatives of* $\underline{x}$ *satisfy*

$$\underline{x}_{ijk} = (\underline{x}_{ij})_k = (\underline{x}_{ik})_j.$$

PROOF. This is immediate from the equality of mixed partials of $\underline{x}_i$. □

The following formula will be used in the proof of the *theorema egregium* of Gauss.

PROPOSITION 10.9.2. *On a surface* M*, we have the relation*

$$\Gamma^q_{i[j,k]} + \Gamma^m_{i[j} \, \Gamma^q_{k]m} = -L_{i[j} L^q{}_{k]}$$

for each set of indices i, j, k, q *(with, as usual, an implied summation over the index* m*).*

PROOF. We leave out the underlines for simplicity. Consider the third partial derivative $x_{ijk} = \frac{\partial^3 x}{\partial u^i \partial u^j \partial u^k}$. Let us calculate its tangential component relative to the basis (x_1, x_2, n) for $\mathbb{R}^3$. Recall that $n_k = L^p{}_k x_p$ and $x_{jk} = \Gamma^\ell_{jk} x_\ell + L_{jk} n$. Thus, we have

$$\begin{aligned}(x_{ij})_k &= \left(\Gamma^m_{ij} x_m + L_{ij} n\right)_k \\ &= \Gamma^m_{ij,k} x_m + \Gamma^m_{ij} x_{mk} + L_{ij} n_k + L_{ij,k} n \\ &= \Gamma^m_{ij,k} x_m + \Gamma^m_{ij}\left(\Gamma^p_{mk} x_p + L_{mk} n\right) + L_{ij}\left(L^p{}_k x_p\right) + L_{ij,k} n.\end{aligned}$$

Grouping together the tangential terms, we obtain

$$\begin{aligned}(x_{ij})_k &= \Gamma^m_{ij,k} x_m + \Gamma^m_{ij}\left(\Gamma^p_{mk} x_p\right) + L_{ij}\left(L^p{}_k x_p\right) + (\ldots)n \\ &= \left(\Gamma^q_{ij,k} + \Gamma^m_{ij}\Gamma^q_{mk} + L_{ij} L^q{}_k\right) x_q + (\ldots)n \\ &= \left(\Gamma^q_{ij,k} + \Gamma^m_{ij}\Gamma^q_{km} + L_{ij} L^q{}_k\right) x_q + (\ldots)n,\end{aligned}$$

since the symbols Γ^q_{km} are symmetric in the two subscripts.

By Lemma 10.9.1, the symmetry in j, k (equality of mixed partials) implies the following identity: $x_{i[jk]} = 0$. Therefore

$$\begin{aligned}0 &= x_{i[jk]} \\ &= (x_{i[j})_{k]} \\ &= \left(\Gamma^q_{i[j,k]} + \Gamma^m_{i[j}\Gamma^q_{k]m} + L_{i[j} L^q{}_{k]}\right) x_q + (\ldots)n.\end{aligned}$$

Since (x_1, x_2, n) is linearly independent, it follows that for each $q = 1, 2$, we have $\Gamma^q_{i[j,k]} + \Gamma^m_{i[j}\Gamma^q_{k]m} + L_{i[j} L^q{}_{k]} = 0$. $\square$

10.10. The *theorema egregium* of Gauss

Recall from Section 9.9 that we have

$$K = \det(L^i{}_j) = 2L^1{}_{[1} L^2{}_{2]} = -\frac{2}{g_{11}} L_{1[1} L^2{}_{2]}. \tag{10.10.1}$$

Theorem 10.10.1 (Theorema egregium). *The Gaussian curvature function $K = K(u^1, u^2)$ of a surface can be expressed in terms of the coefficients of the first fundamental form alone (and their first and second derivatives) as follows:*

$$K = \frac{2}{g_{11}}\left(\Gamma^2_{1[1,2]} + \Gamma^j_{1[1}\Gamma^2_{2]j}\right), \tag{10.10.2}$$

where the symbols Γ^k_{ij} can be expressed in terms of the derivatives of g_{ij} be the formula $\Gamma^k_{ij} = \frac{1}{2}(g_{i\ell,j} - g_{ij,\ell} + g_{j\ell,i})g^{\ell k}$, where (g^{ij}) is the inverse matrix of (g_{ij}).

In other words, we have

$$K = \frac{1}{g_{11}}\left(\Gamma_{11,2}^2 - \Gamma_{12,1}^2 + \Gamma_{11}^1\Gamma_{21}^2 - \Gamma_{12}^1\Gamma_{11}^2 + \Gamma_{11}^2\Gamma_{22}^2 - \Gamma_{12}^2\Gamma_{12}^2\right) \tag{10.10.3}$$

COROLLARY 10.10.2. *If the first fundamental form is diagonal, then*

$$K = g^{11}\left(\Gamma_{11,2}^2 - \Gamma_{12,1}^2 + \Gamma_{11}^1\Gamma_{21}^2 - \Gamma_{12}^1\Gamma_{11}^2 + \Gamma_{11}^2\Gamma_{22}^2 - \Gamma_{12}^2\Gamma_{12}^2\right)$$

PROOF. If the matrix of the first fundamental form is diagonal then one can make the substitution $g^{11} = \frac{1}{g_{11}}$ in (10.10.3). □

PROOF OF *theorema egregium*. We present a streamlined version of do Carmo's proof [**Ca76**, p. 233]. The proof is in 3 steps.

(1) We express the third partial derivative x_{ijk} in terms of both the Γ's (intrinsic information) and the L's (extrinsic information).
(2) The equality of mixed partials yields an identification of a suitable expression in terms of the Γ's, with a certain combination of the L's.
(3) The combination of the L's is expressed in terms of Gaussian curvature.

The first two steps were carried out in Proposition 10.9.2. We choose the valus $i = j = 1$ and $k = q = 2$ for the indices. Applying Theorem 9.9.1(c) (namely, identity (10.10.1)) we obtain

$$\begin{aligned}\Gamma_{1[1,2]}^2 + \Gamma_{1[1}^m\Gamma_{2]m}^2 &= -L_{1[1}L^2{}_{2]} \\ &= g_{1i}L^i{}_{[1}L^2{}_{2]} \\ &= g_{11}L^1{}_{[1}L^2{}_{2]}\end{aligned}$$

since the term $L^2{}_{[1}L^2{}_{2]} = 0$ vanishes. This yields the desired formula for K and completes the proof of the *theorema egregium*. □

CHAPTER 11

Laplace–Beltrami, Gauss–Bonnet

11.1. The Laplace–Beltrami operator

Let M be a surface. Recall that a (u^1, u^2)-chart in which the metric on M becomes conformal (see Definition 10.1.1)[1] to the standard flat metric, is referred to as *isothermal coordinates*. The existence of the latter is proved in [**Bes87**].

Consider a metric on M of the form $g_{ij} = \lambda(u^1, u^2)\delta_{ij}$, where $\lambda > 0$. Then (u^1, u^2) are isothermal coordinates.

Definition 11.1.1. The *Laplace–Beltrami operator* of the metric $\lambda(u^1, u^2)\delta_{ij}$ is the operator

$$\Delta_{LB} = \frac{1}{\lambda}\left(\frac{\partial^2}{\partial(u^1)^2} + \frac{\partial^2}{\partial(u^2)^2}\right).$$

The notation *means* that when we apply the operator Δ_{LB} to a function $h = h(u^1, u^2)$, we obtain $\Delta_{LB}(h) = \frac{1}{\lambda}\left(\frac{\partial^2 h}{\partial(u^1)^2} + \frac{\partial^2 h}{\partial(u^2)^2}\right)$.

In more readable form, for a metric $g_{ij} = \lambda(x, y)\delta_{ij}$ and an arbitrary function $h = h(x, y) \in C^2(\mathbb{R}^2)$, we have

$$\Delta_{LB}(h) = \frac{1}{\lambda}\left(\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2}\right).$$

In other words $\Delta_{LB} h = \frac{1}{\lambda}\Delta_0 h$ where Δ_0 is the flat Laplacian with respect to coordinates (x, y).

11.2. Laplace-Beltrami formula for Gaussian curvature

Let $\ln x$ be the natural logarithm so that $\frac{d}{dx}\ln x = \frac{1}{x}$.

Theorem 11.2.1. *Given a metric in isothermal coordinates with metric coefficients $g_{ij} = \lambda(u^1, u^2)\delta_{ij}$, its Gaussian curvature is minus half the Laplace-Beltrami operator applied to the* ln *of the conformal factor λ:*

$$K = -\frac{1}{2}\,\Delta_{LB}(\ln \lambda). \tag{11.2.1}$$

[1]See also Definition 16.10.1

Γ_{ij}^1	$j=1$	$j=2$
$i=1$	μ_1	μ_2
$i=2$	μ_2	$-\mu_1$

Γ_{ij}^2	$j=1$	$j=2$
$i=1$	$-\mu_2$	μ_1
$i=2$	μ_1	μ_2

TABLE 11.2.1. Symbols Γ_{ij}^k of a metric $e^{2\mu(u^1,u^2)}\delta_{ij}$

PROOF. In Section 10.1 we computed the Gamma symbols with respect to isothermal coordinates. Recall from (10.10.2) that

$$K = \frac{2}{g_{11}}\left(\Gamma_{1[1,2]}^2 + \Gamma_{1[1}^j\Gamma_{2]j}^2\right), \tag{11.2.2}$$

Step 1. We write $\lambda = e^{2\mu}$ and tabulate the Γ in Table 11.2.1. We have from Table 11.2.1 (based on Table 10.1.1):

$$2\Gamma_{1[1,2]}^2 = \Gamma_{11,2}^2 - \Gamma_{12,1}^2 = -\mu_{22} - \mu_{11}.$$

Step 2. The $\Gamma\Gamma$ term in the expression (11.2.2) for the Gaussian curvature vanishes:

$$\begin{aligned}
2\Gamma_{1[1}^j\Gamma_{2]j}^2 &= 2\Gamma_{1[1}^1\Gamma_{2]1}^2 + 2\Gamma_{1[1}^2\Gamma_{2]2}^2 \\
&= \Gamma_{11}^1\Gamma_{21}^2 - \Gamma_{12}^1\Gamma_{11}^2 + \Gamma_{11}^2\Gamma_{22}^2 - \Gamma_{12}^2\Gamma_{12}^2 \\
&= \mu_1\mu_1 - \mu_2(-\mu_2) + (-\mu_2)\mu_2 - \mu_1\mu_1 \\
&= 0.
\end{aligned}$$

Step 3. Since the $\Gamma\Gamma$ term vanishes by Step 2, from formula (10.10.2) we have

$$K = \frac{2}{\lambda}\Gamma_{1[1,2]}^2 = -\frac{1}{\lambda}(\mu_{11} + \mu_{22}) = -\Delta_{LB}\mu.$$

Meanwhile, $\Delta_{LB}\ln\lambda = \Delta_{LB}(2\mu) = 2\Delta_{LB}(\mu)$, proving the result. $\square$

Setting $\lambda = f^2$, we can restate the theorem as follows.

COROLLARY 11.2.2. *Given a metric in isothermal coordinates with metric coefficients $g_{ij} = f^2(u^1,u^2)\delta_{ij}$, its Gaussian curvature is minus the Laplace-Beltrami operator of the* ln *of the conformal factor f:*

$$K = -\Delta_{LB}(\ln f). \tag{11.2.3}$$

Either one of the formulas (10.10.2), (11.2.1), or (11.2.3) can serve as the intrinsic definition of Gaussian curvature, replacing the extrinsic definition (10.10.1); *cf.* Remark 10.7.

11.3. Area elements of the surface, orientability

Consider a surface $M \subseteq \mathbb{R}^3$ with local parametrisation $\underline{x}(u^1, u^2)$. We have

$$\sqrt{\det(g_{ij})} = \left|x_1 \times x_2\right|,$$

where $x_i = \frac{\partial \underline{x}}{\partial u^i}$.[2] We saw two cases of area elements of surfaces:

(1) the case of the area element $r\, dr d\theta$ in the plane in polar coordinates as in Section 5;
(2) the spherical element of area $\sin\varphi\, d\varphi\, d\theta$.

More generally, we have the following definition, which already appeared in Definition 8.6.1.

Definition 11.3.1. The *area element* dA_M of the surface M is

$$dA_M = \sqrt{\det(g_{ij})}\, du^1 du^2 = |x_1 \times x_2|\, du^1 du^2 \tag{11.3.1}$$

where the g_{ij} are the metric coefficients of the surface with respect to the parametrisation $\underline{x}(u^1, u^2)$.

We have used the subscript M so as to specify which surface we are dealing with.[3]

Definition 11.3.2. A surface $M \subseteq \mathbb{R}^3$ is *orientable* if it admits a continuous unit normal vector field $N = N_p$, defined at each point $p \in M$.

The vector field N is a globally defined field along an orientable surface. Note that the image vector N can be thought of as an element of the unit sphere: $N_p \in S^2 \subseteq \mathbb{R}^3$. This observation leads to the definition of the Gauss map for surfaces in Section 11.4.

11.4. Gauss map for surfaces in $\mathbb{R}^3$

Let $M \subseteq \mathbb{R}^3$ be an orientable surface. Let N be a globally defined normal vector field along M.

Remark 11.4.1. Each point $p \in M$ lies in a neighborhood with a suitable parametrisation $\underline{x}(u^1, u^2)$. At a point $p = \underline{x}(u^1, u^2)$, we have a normal vector $n(u^1, u^2)$, obtained by normalizing the vector product $x_1 \times x_2$, which coincides with the globally selected normal N. Thus, we have

$$n(u^1, u^2) = N_{\underline{x}(u^1, u^2)}.$$

[2]Cf. the Binet–Cauchy identity 11.10.3.

[3]This notion of area is discussed in more detail in Section 16.9.

Definition 11.4.2. The *Gauss map*

$$G_M\colon M \to S^2$$

of an orientable surface $M \subseteq \mathbb{R}^3$ is the map sending each point $p = \underline{x}(u^1, u^2)$ to the normal vector $G(p) = n(u^1, u^2)$, where at every point we choose the normal that coincides with the global vector field N at p.

Example 11.4.3 (Gauss map of the plane is constant). If $M = \mathbb{R}^2$ is the xy-plane then the normal vector at every point is e_3 and therefore $G_M(p) = e_3$ for all points $p \in M$. Thus the Gauss map is a constant map in this case.

Example 11.4.4. If $M = S^2_r$ is the sphere of radius $r > 0$ then we have $G_M(p) = \frac{1}{r}p$ for all $p \in M$. Thus the Gauss map G_M is proportional to the identity map in this case: $G_{S^2} = \frac{1}{r}\,\mathrm{Id}_{S^2}$.

Recall that $W_p : T_p \to T_p$ is the Weingarten map.

Lemma 11.4.5. *If $K \neq 0$ at a point $p \in M$, then the tangent vectors $W_p(x_1), W_p(x_2) \in T_p$ are linearly independent.*

Proof. The coefficients of the matrix $(L^i{}_j)$ are by definition the coordinates of $\frac{\partial n}{\partial u^1}, \frac{\partial n}{\partial u^2}$ with respect to the basis (x_1, x_2). Hence independence of the vectors n_1 and n_2 is equivalent to the condition $K = \det(L^i{}_j) \neq 0$. $\square$

Theorem 11.4.6. *Choose a point $p = \underline{x}(a, b) \in M$ of the surface. If the Gaussian curvature is nonzero at p, then the map $n(u^1, u^2)$ from (a neighborhood of) M to S^2 produces a regular parametrisation of a spherical neighborhood of the point*

$$G_M(p) = n(a, b) \in S^2$$

on the sphere, where G_M is the Gauss map.

Proof. The parametrisation is given by the map G_M to the sphere. We have

$$\frac{\partial}{\partial u^i}(G_M) = \frac{\partial}{\partial u^i}(n(u^1, u^2)) = \frac{\partial n}{\partial u^i} = W_p(x_i).$$

By Lemma 11.4.5, the two partial derivatives of the Gauss map G_M are linearly independent. Hence the parametrisation defined by $n(u^1, u^2)$ is regular in a neighborhood of p. $\square$

11.5. Sphere S^2 parametrized by Gauss map of M

The vector $n(u^1, u^2)$ was originally defined as a normal to the original surface $M \subseteq \mathbb{R}^3$. We now with to think of it as giving a parametrisation of an open neighborhood on the unit sphere S^2 via the Gauss map from M to S^2.

DEFINITION 11.5.1 (Coefficients $\tilde{g}_{\alpha\beta}$). Assume $K_p \neq 0$ at a point $p \in M$, and consider a neighborhood of the point $G_M(p)$ in S^2. Let

$$\tilde{g}_{\alpha\beta} = n_\alpha \cdot n_\beta$$

be the metric coefficients of the regular parametrisation $n(u^1, u^2)$ in the neighborhood, where $n_\alpha = \frac{\partial n(u^1,u^2)}{\partial u^\alpha}$ for all $\alpha = 1, 2$, $\beta = 1, 2$.

Now consider the area element dA_{S^2} (defined in Section 11.3) of the unit sphere.

THEOREM 11.5.2. *Let $M \subseteq \mathbb{R}^3$ be a surface, and consider the corresponding parametrisation $n(u^1, u^2)$ of a neighborhood on the sphere S^2.*[4] *Then the area element dA_{S^2} of S^2 can be expressed as*

$$dA_{S^2} = \sqrt{\det(\tilde{g}_{\alpha\beta})}\, du^1 du^2 = |n_1 \times n_2| du^1 du^2. \tag{11.5.1}$$

PROOF. Formula (11.5.1) is the definition of the element of area for the surface S^2, applied to the chosen parametrisation $n(u^1, u^2)$ as defined in Theorem 11.4.6, in place of the standard parametrisation in spherical coordinates.[5] □

REMARK 11.5.3. We have used the subscript S^2 to distinguish the area element dA_{S^2} of the sphere S^2 from the area element dA_M of the surface M as in (11.3.1). Note that we need to distinguish the two area elements

$$dA_{S^2} \quad \text{and} \quad dA_M$$

because both will occur in the proof of the Gauss–Bonnet theorem in Section 11.7.

11.6. Comparison of two parametrisations

When the Gaussian curvature is nonzero, the normal vector of a surface $M \subseteq \mathbb{R}^3$ allows us to define local coordinates $n(u^1, u^2)$ on the unit sphere S^2 as in Section 11.4.

PROPOSITION 11.6.1. *Consider the following data:*

(1) *the metric coefficients (g_{ij}) of the parametrisation $\underline{x}(u^1, u^2)$ of the surface M, and*
(2) *the metric coefficients $(\tilde{g}_{\alpha\beta})$ of the sphere S^2 relative to the parametrisation $n(u^1, u^2)$.*

[4]As in Theorem 11.4.6.
[5]Cf. Binet–Cauchy in note 14.

Then we have the identity

$$\sqrt{\det(\tilde{g}_{\alpha\beta})} = |K_M(u^1, u^2)|\sqrt{\det(g_{ij})}$$

where $K = K(u^1, u^2)$ *is the Gaussian curvature of the surface* M*.*

Proof. Let $L = (L^i{}_j)$ be the matrix of the Weingarten map W_p with respect to the basis (x_1, x_2) at a point $p = \underline{x}(u^1, u^2)$. By definition of curvature we have $K = \det(L)$. Recall that the coefficients $L^i{}_j$ of the Weingarten map are defined via the decomposition

$$n_\alpha = x_i\, L^i{}_\alpha. \tag{11.6.1}$$

Consider the 3×2-matrices $A = [x_1 \quad x_2]$ and $B = [n_1 \quad n_2]$. Then formula (11.6.1) implies by definition of matrix multiplication that

$$B = AL.$$

Therefore the corresponding Gram matrices satisfy

$$\mathrm{Gram}(n_1, n_2) = B^t\, B = (AL)^t AL = L^t(A^t A)L = L^t\ \mathrm{Gram}(x_1, x_2)\ L.$$

Applying the determinant, we obtain

$$\det\left(\mathrm{Gram}(n_1, n_2)\right) = \det\left(\mathrm{Gram}(x_1, x_2)\right) \det{}^2(L),$$

completing the proof of the lemma since $\det(L) = K$, and the Gram matrix represents the first fundamental form. [6] □

11.7. Gauss–Bonnet theorem for surfaces with $K_M > 0$

The Gauss–Bonnet theorem for surfaces[7] asserts that an integral of curvature has topological significance.[8] We will only treat the Gauss–Bonnet theorem in the case $K > 0$.

Example 11.7.1. A typical example of a compact surface of positive Gaussian curvature is an ellipsoid.

[6] By the Cauchy-Binet formula, the desired identity is equivalent to the formula $|n_{u^1} \times n_{u^2}| = |\det(L^i{}_j)|\ |x_{u^1} \times x_{u^2}|$, immediate from the observation that a linear map multiplies the area of parallelograms by its determinant. Namely, the Weingarten map sends each vector x_i to n_i: $W_p(x_i) = n_i \in T_p$. A stronger identity is true that is sensitive to the sign of the Gaussian curvature: $W(u)\times W(v) = \det(W)(u\times v)$ where (u, v) is any basis of the tangent plane; e.g., the basis (x_1, x_2).

[7] as well as the theorem on the total curvature of a plane curve (Theorem 12.5.4).

[8] See also Section 12.8 for a version of the theorem for plane domains.

THEOREM 11.7.2 (Special case of Gauss–Bonnet). *Let M be an orientable compact surface in $\mathbb{R}^3$ of positive Gaussian curvature. Then the curvature integral satisfies*

$$\int_M K_p \, dA_M = 4\pi,$$

where K_p is the Gaussian curvature of M at the point p.

REMARK 11.7.3. Note that $4\pi = 2\pi\,\chi(S^2)$ where χ is the Euler characteristic (see Section 11.10).

To prove Theorem 11.7.2, we will use the existence of global coordinates[9] on the sphere to write down a concrete version of the calculation. We recall the following five items.

(1) $\underline{x}(u^1, u^2)$ is a regular parametrisation of a surface $M \subseteq \mathbb{R}^3$.
(2) At every point of M where the parametrisation $\underline{x}$ is defined, we have the normal vector $n(u^1, u^2)$ so that (x_1, x_2, n) is a basis for $\mathbb{R}^3$.
(3) If M is an orientable surface, a continuous Gauss map
$$G_M \colon M \to S^2$$
is defined by the normal vector $n(u^1, u^2)$ as in Section 11.4.
(4) (θ, φ) are the usual spherical coordinates on the unit sphere.
(5) If $K_p \neq 0$ then the Gauss map $G_M \colon M \to S^2$ is invertible near p.

Since the Gauss map is invertible, we can consider the inverse map

$$G_M^{-1} \colon S^2 \to M. \tag{11.7.1}$$

DEFINITION 11.7.4. Let $I_\theta = [0, 2\pi]$ and $I_\varphi = [0, \pi]$. We consider the usual parametrisation

$$\sigma = \sigma(\theta, \phi) \colon I_\theta \times I_\varphi \to S^2$$

given by

$$\sigma(\theta, \varphi) = (\sin\varphi\cos\theta, \sin\varphi\sin\theta, \cos\varphi) \in S^2,$$

of the sphere as a surface of revolution (omitting the problematic poles), as in Section 5.4.

We also construct a special parametrisation of M based on the parametrisation σ of the sphere, as follows.

[9]These have singularities only at the north and south poles, which does not affect the calculation of area.

Definition 11.7.5. Let $\underline{x}(\theta, \varphi)$ be the parametrisation of M given by the composition $\underline{x} = G_M^{-1} \circ \sigma$ using (11.7.1):

$$I_\theta \times I_\varphi \xrightarrow{\sigma} S^2 \xrightarrow{G^{-1}} M,$$

and denote by (g_{ij}) the corresponding metric coefficients.

In Section 11.8, we will use this parametrisation to compute the curvature integral on M.

11.8. Proof of Gauss–Bonnet Theorem

We can dispense with *local* coordinate charts and instead compute the Gauss–Bonnet integral over M relative to the global coordinates (θ, φ) as in Definition 11.7.5.[10] Let (g_{ij}) be the metric coefficients of M with respect to the parametrisation of Definition 11.7.5. Then the area element of M can be written as

$$dA_M = \sqrt{\det(g_{ij})}\, d\theta d\varphi.$$

We will exploit the relation

$$K(\theta, \varphi)\sqrt{\det(g_{ij})} = \sqrt{\det(\tilde{g}_{\alpha\beta})} \tag{11.8.1}$$

from Section 11.6, where $u^1 = \theta$ and $u^2 = \varphi$. Here $\tilde{g}_{\alpha\beta}$ are the metric coefficients of the standard metric on S^2 with respect to parametrisation defined by the Gauss map G of M. We write the double integral $\iint$ as shorthand for the iterated integral

$$\int_{\varphi=0}^{\varphi=\pi} \int_{\theta=0}^{\theta=2\pi}.$$

Then we obtain using identity (11.8.1):

$$\begin{aligned}
\int_M K(\theta, \varphi)\, dA_M &= \iint K(\theta, \varphi)\sqrt{\det(g_{ij})}\, d\theta d\varphi \\
&= \iint \sqrt{\det(\tilde{g}_{\alpha\beta})}\, d\theta d\varphi \\
&= \iint dA_{S^2} \\
&= \text{area}(S^2) = 4\pi,
\end{aligned}$$

as required.[11]

[10] These coordinates only omit two points, which does not affect area calculations.

[11] **Alternative proof**. This material is optional. Here we present another proof of Gauss–Bonnet theorem. The convexity of the surface guarantees that

11.9. Gauss–Bonnet with boundary

There are various generalisations of the Gauss–Bonnet theorem of Section 11.8. One of them is the following version with boundary.

THEOREM 11.9.1 (Version with boundary of Gauss–Bonnet theorem). *Let M be a surface. Consider a geodesic triangle (homeomorphic to a disk) $T \subseteq M$ with angles α, β, γ. Then the integral of the Gaussian curvature over T is the angular excess:*

$$\int_T K dA = \alpha + \beta + \gamma - \pi.$$

COROLLARY 11.9.2. *The area of a spherical triangle with angles α, β, and γ is the angular excess $\alpha + \beta + \gamma - \pi$.*

PROOF. We apply the local Gauss–Bonnet theorem and notice that for $M = S^2$, we have $K = 1$ at every point. $\square$

REMARK 11.9.3 (Historical remarks). Stillwell [**6**, p. 329] notes that the result for spherical triangles was known already to Thomas Harriot in 1603. Giusti reports that Cavalieri found the area of a spherical triangle to be the surplus of the sum of the angles over π (when the radius is normalized), which is an important special case of the Gauss-Bonnet theorem [**3**, p. 13].

11.10. Euler characteristic and Gauss–Bonnet theorem

DEFINITION 11.10.1. Assume a surface M is partitioned into triangles. Then the Euler characteristic $\chi(M)$ of M is

$$\chi(M) = V - E + F,$$

where V, E, F are respectively the numbers of vertices, edges, and faces (i.e., triangles) of M.

The Euler characteristic of a closed embedded surface in Euclidean 3-space can be computed via the integral of the Gaussian curvature.[12]

---

the map n is one-to-one (compare with the proof of Theorem on closed curves in Section 12.4). The integrand $K\, dA_M$ in a coordinate chart (u^1, u^2) can be written as $K(u^1, u^2)\, dA_M$. By Proposition 11.6.1, we have $K(u^1, u^2)\, dA_M = K(u^1, u^2)\sqrt{\det(g_{ij})} du^1 du^2 = \sqrt{\det(\tilde{g}_{\alpha\beta})} du^1 du^2 = dA_{S^2}$. Thus, the expression $K\, dA_M$ coincides with the area element dA_{S^2} of the unit sphere S^2 in every coordinate chart. Hence we can write $\int_M K\, dA_M = \int_{S^2} dA_{S^2} = 4\pi$, proving the theorem.

[12] It can be thought of as a generalisation of the rotation index of a plane closed curve; see Chapter 12.

Theorem 11.10.2 (Gauss–Bonnet). *The Euler characteristic* $\chi(M)$ *of a compact surface* M *satisfies*

$$\int_M K_p \, dA_M = 2\pi\, \chi(M), \tag{11.10.1}$$

where K_p *is the Gaussian curvature at the point* $p \in M$*.*

Here we no longer assume that the curvature is positive.[13]

One way of proving Gauss–Bonnet for embedded surfaces is to use the notion of algebraic degree for maps between surfaces, similar to the algebraic degree of a self-map of a circle.

We note the following.

(1) The Gauss–Bonnet theorem holds for all surfaces, whether orientable or not.
(2) For the real projective plane $\mathbb{RP}^2$ we have $\chi(\mathbb{RP}^2) = 1$. Therefore for every metric on $\mathbb{RP}^2$ we have $\int_{\mathbb{RP}^2} K\, dA_{\mathbb{RP}^2} = 2\pi$.
(3) For the torus T^2 we have $\chi(T^2) = 0$. Therefore for every metric on T^2 we have $\int_{T^2} K\, dA_{T^2} = 0$. It follows that the torus does not admit any metric of positive Gaussian curvature.[14]

---

[13] The relation (11.10.1) is similar to the line integral expression for the rotation index in formula (13.1.1).

[14] This material is optional. For a reader familiar with elements of Riemannian geometry including Jacobi fields, it is worth mentioning that the Jacobi equation $y'' + Ky = 0$ of a Jacobi field y on M (expressing an infinitesimal variation by geodesics) sheds light on the nature of curvature in a way that no mere formula for K could. Thus, in positive curvature, geodesics converge, while in negative curvature, they diverge. However, to prove the Jacobi equation, one would need to have already an intrinsically well-defined quantity on the left hand side, $y'' + Ky$, of the Jacobi equation. In particular, one would need an already intrinsic notion of curvature K. Thus, a proof of the *theorema egregium* necessarily precedes the deeper insights provided by the Jacobi equation. Similarly, the Gaussian curvature at $p \in M$ is the first significant term in the asymptotic expansion of the length of a "small" circle of center p. This fact, too, sheds much light on the nature of Gaussian curvature. However, to define what one means by a "small" circle, requires introducing higher order notions such as the exponential map, which are usually understood at a later stage than the notion of Gaussian curvature, *cf.* [**Ca76, Car92, Ch93, GaHL04**].

Theorem 11.10.3 (Binet–Cauchy identity). *The 3-dimensional case of the Binet–Cauchy identity is the identity* $(a\cdot c)(b\cdot d) = (a\cdot d)(b\cdot c) + (a\times b)\cdot(c\times d)$*, where* a*,* b*,* c*, and* d *are vectors in* $\mathbb{R}^3$*.*

The formula can also be written as a formula giving the dot product of two wedge products, namely $(a\times b)\cdot(c\times d) = (a\cdot c)(b\cdot d) - (a\cdot d)(b\cdot c)$. A special case of Binet–Cauchy is the case of vectors $a = c$ and $b = d$, when we obtain $|a\times b|^2 = |a|^2|b|^2 - |a\cdot b|^2$. When both vectors a, b are unit vectors, we obtain the usual relation $1 = \cos^2(\phi) + \sin^2(\phi)$ (here the vector product gives the sine and the scalar product gives the cosine) where ϕ is the angle between the vectors. Let $a = x_1$ and $b = x_2$ be the two tangent vectors to the surface with parametrisation $x(u^1, u^2)$. Then $|x_1 \times x_2|^2 = g_{11}g_{22} - g_{12}^2 = \det(g_{ij})$. Equivalently, we have $|x_1 \times x_2| = \sqrt{\det(g_{ij})}$. This can be thought of as the area of the parallelogram spanned by the two vectors.

CHAPTER 12

Signed curvature of curves; total curvature

12.1. Angle $\theta(s)$

The signed curvature of a plane curve is a refinement of the ordinary curvature of curves. Part of the motivation for introducing the refinement is that the classical definition of Gaussian curvature is in terms of the product of signed curvatures (see Section 12.10). Signed curvatures are defined below.

We identify $\mathbb{R}^2$ with $\mathbb{C}$ so that a vector in the plane can be written as a complex number. Let s be an arclength parameter along a curve $\alpha(s)$ with tangent vector $v(s) = \alpha'(s)$. Since v is of unit norm we can write it as $v = e^{i\theta}$.

Definition 12.1.1 (function theta). The angle $\theta(s)$ along the curve $\alpha(s)$ with tangent vector $v(s) = \alpha'(s)$ is defined in one of the following two equivalent ways:

(1) We write $v(s) = e^{i\theta(s)}$, where the angle $\theta(s)$ is measured counterclockwise, from the positive ray of the x-axis, to the vector $v(s)$.
(2) Using a suitable branch of the complex logarithm, we can also express $\theta(s)$ as follows: $\theta(s) = \frac{1}{i}\log v(s) = -i\log v(s)$.

Here log is a suitable branch of the inverse of the complex exponential function e^z, not to be confused with the real function ln.

Lemma 12.1.2. *If* $\alpha(s) = (x(s), y(s))$ *then* $\frac{dx}{ds} = \cos\theta$ *and* $\frac{dy}{ds} = \sin\theta$.

Proof. By definition we have $v(s) = \frac{dx}{ds} + i\,\frac{dy}{ds}$. Since $v(s) = e^{i\theta} = \cos\theta + i\sin\theta$, we obtain the formulas of the lemma. □

Lemma 12.1.3. *We have the relation* $\frac{d}{d\theta}e^{i\theta} = ie^{i\theta}$, *where* $\theta(s)$ *is the angle of Definition* 12.1.1.

This was proved in complex functions.

12.2. Signed curvature $\tilde{k}$ as $\theta'(s)$

We are now ready to define a refinement of the curvature function $k_\alpha(s)$, called the signed curvature.

DEFINITION 12.2.1. The *signed curvature* function $\tilde{k}_\alpha$ of the plane curve α is

$$\tilde{k}_\alpha(s) = \frac{d\theta}{ds}. \tag{12.2.1}$$

This is discussed in more detail in Section 12.7. We have the following relation between the signed curvature and the ordinary curvature of a curve as defined in Section 4.2.

THEOREM 12.2.2. *We have* $\left|\tilde{k}_\alpha\right| = k_\alpha$.

PROOF. We differentiate $v(s) = e^{i\theta(s)}$ by chain rule to obtain

$$\frac{dv}{ds} = ie^{i\theta(s)}\frac{d\theta}{ds},$$

and therefore

$$|\tilde{k}_\alpha(s)| = \left|\frac{d\theta}{ds}\right| = \left|\frac{dv}{ds}\right| = \left|\frac{d^2\alpha}{ds^2}\right| = k_\alpha(s)$$

as required. □

12.3. Signed curvature with respect to arbitrary parameter

We start with the following data.

(1) $\alpha(s)$ an arclength parametrisation of a plane curve;
(2) $k_\alpha(s) = |\alpha''(s)|$ is the curvature of the curve (see formula (4.2.1));
(3) $\theta(s)$ is the angle defined so that $\alpha'(s) = e^{i\theta(s)}$ in the plane;
(4) $\tilde{k}_\alpha(s) = \frac{d\theta}{ds}$ is the signed curvature (see Section 12.1).

The signed curvature of a plane curve can be expressed in terms of an arbitrary parameter t of a regular parametrisation, as follows. We use dots (Newton's notation) for derivatives with respect to t.

THEOREM 12.3.1. *Let $\alpha(t) = (x(t), y(t))$ be an arbitrary regular parametrisation (not necessarily arclength) of a plane curve. Then signed curvature satisfies the following equivalent formulas:*

$$\tilde{k}_\alpha(t) = \begin{cases} \dfrac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{(\dot{x}^2 + \dot{y}^2)^{3/2}} \\ \dfrac{\det\begin{pmatrix}\dot{\alpha} & \ddot{\alpha}\end{pmatrix}}{|\dot{\alpha}|^3} \ \textit{for the 2 by 2 matrix} \begin{pmatrix}\dot{\alpha} & \ddot{\alpha}\end{pmatrix}. \end{cases} \tag{12.3.1}$$

PROOF. With respect to an arbitrary parameter t, the components of $\dot{\alpha}(t)$ are $\dot{x}(t)$ and $\dot{y}(t)$. Hence we have $\tan\theta = \frac{\dot{y}}{\dot{x}}$ or equivalently

$$\dot{x}\sin\theta = \dot{y}\cos\theta. \tag{12.3.2}$$

Differentiating (12.3.2) with respect to t, we obtain by product rule and chain rule

$$\ddot{x}\sin\theta + \dot{x}\cos\theta\,\frac{d\theta}{dt} = \ddot{y}\cos\theta - \dot{y}\sin\theta\,\frac{d\theta}{dt}. \tag{12.3.3}$$

Grouping together terms containing $\frac{d\theta}{dt}$, we obtain

$$\frac{d\theta}{dt}(\dot{x}\cos\theta + \dot{y}\sin\theta) = -\ddot{x}\sin\theta + \ddot{y}\cos\theta. \tag{12.3.4}$$

The arclength parameter s is defined by $s(t) = \int_0^t |\dot{\alpha}|$ so that $\frac{ds}{dt} = |\dot{\alpha}|$. Multiplying (12.3.4) by $\frac{ds}{dt}$, we obtain

$$\frac{d\theta}{dt}\left(\dot{x}\cos\theta\frac{ds}{dt} + \dot{y}\sin\theta\frac{ds}{dt}\right) = -\ddot{x}\sin\theta\frac{ds}{dt} + \ddot{y}\cos\theta\frac{ds}{dt}.$$

Recall that by Lemma 12.1.2 we have

$$\begin{cases} \cos\theta = \frac{dx}{ds}, \\ \sin\theta = \frac{dy}{ds}. \end{cases}$$

Therefore by chain rule, we obtain $\frac{d\theta}{dt}\left(\dot{x}^2 + \dot{y}^2\right) = -\ddot{x}\dot{y} + \ddot{y}\dot{x}$ so that

$$\frac{d\theta}{dt} = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{\dot{x}^2 + \dot{y}^2}.$$

Furthermore, $\frac{d\theta}{dt} = \frac{d\theta}{ds}\frac{ds}{dt}$ and therefore $\frac{d\theta}{ds} = \frac{\dot{x}\ddot{y}-\ddot{x}\dot{y}}{\left(\frac{ds}{dt}\right)^3}$ since $\frac{ds}{dt} = |\dot{\alpha}(t)| = |\frac{d\alpha}{dt}|$. Thus

$$\frac{d\theta}{ds} = \frac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{|\dot{\alpha}|^3}$$

and by (12.2.1) we obtain the desired pair of formulas (12.3.1) for the signed curvature. □

Example 12.3.2. Calculate the curvature of the graph of $y = f(x)$ at a critical point $x = c$ using formula (12.3.1).

We will return to the study of signed curvature in Section 12.7.

12.4. Global geometry of Jordan curves

In this section, we begin the study of the *global* geometry of plane curves. The curvature invariants we have studied thus far are local invariants. The global geometry of surfaces is studied in Section 11.3.

Definition 12.4.1. A *Jordan curve* in the Euclidean plane $\mathbb{R}^2 = \mathbb{C}$ is a non-selfintersecting closed curve that can be represented by a continuous map $\alpha(s)\colon [0, L] \to \mathbb{R}^2$ such that $\alpha(0) = \alpha(L)$ and α is injective elsewhere.

THEOREM 12.4.2 (Jordan curve theorem). *A Jordan curve separates the plane into two open regions: a bounded region and an unbounded region.*

The bounded region is called the *interior* region. We will only deal with smooth (i.e., infinitely differentiable) regular maps α. Consider a smooth Jordan curve $J \subseteq \mathbb{C}$ of length L. By Theorem 4.15.3 there is an arclength parametrisation $\alpha(s)$ of J, where

1. $s \in [0, L]$;
2. $\alpha(0) = \alpha(L)$;
3. $\alpha'(0) = \alpha'(L)$;
4. the tangent vector $v(s) = \alpha'(s)$ satisfies $v(s) \in S^1 \subseteq \mathbb{C}$ where S^1 is the unit circle.
5. $v(s) = \alpha'(s) = e^{i\theta(s)}$ where $\theta \in [0, L]$.

See also note.[1]

DEFINITION 12.4.3. A smooth Jordan curve $J \subseteq \mathbb{R}^2$ is called *strictly convex*[2] if $\tilde{k} > 0$.

DEFINITION 12.4.4 (Orientation). A *counterclockwise parametrisation* (orientation) of a strictly convex Jordan curve J is the arclength parametrisation such that the angle

$$\theta(s) \colon [0, L] \to [0, 2\pi]$$

is an *increasing* function on $[0, L]$.

See also note.[3]

---

[1]This is mainly motivation for the Gauss–Bonnet theorem. The normal vector $n(s) = iv(s)$ defines a map G to the unit circle called the *Gauss map*, as follows. The *Gauss map* G of a smooth Jordan curve $J \subseteq \mathbb{C}$ is the map $G \colon J \to S^1, \quad \alpha(s) \mapsto n(s) = iv(s)$ for each point $\alpha(s) \in J$, where $v(s) = \alpha'(s)$. The Gauss map is usually defined (as we did) using the normal vector $n(s)$ (see note 7) though (in the case of curves unlike the case of surfaces) one could use $v(s)$ as well. Example: If the curve J is a circle C_r of radius r centered at the origin, then its Gauss map is the map $G = \frac{1}{r}\mathrm{Id}_J$, i.e., a multiple of the identity map of the circle.

[2]kamur with kuf

[3]Recall the following set-theoretic notions. The set-theoretic complement of a subset $B \subseteq A$ is denoted $A \setminus B$. Any line $\ell \subseteq \mathbb{R}^2$ divides the plane into two open halfplanes, namely the two connected components (rechivei kshirut) of the set complement $\mathbb{R}^2 \setminus \ell$. More generally, one could define a strictly convex curve by requiring one of the following two equivalent conditions: (1) the interior of each segment joining a pair of points of J is contained in the interior region (Tchum pnimi) of J; (2) consider the tangent line T_p to J at a point $p \in J$; then the complement $J \setminus \{p\}$ lies entirely in one of the open halfplanes of $\mathbb{R}^2 \setminus T_p$ defined by the tangent line, for all $p \in J$.

See further in note.[4]

---

[4]Theorem. Given a strictly convex smooth regular Jordan curve $J \subseteq \mathbb{C}$, the Gauss map $G\colon J \to S^1$ is one-to-one and onto. [Proof of the "onto" part] To fix ideas we will assume that J is parametrized counterclockwise. We will work instead with the map $v\colon J \to S^1$ defined by the tangent vector v. This is legitimate since the tangent vector only differs from the normal vector by a $90°$ rotation. Let $\alpha(s)$ be an arclength parametrisation of J. First we consider the case of "horizontal" vectors $v(s)$ in the (x, y)-plane. These occur at points of $J \subseteq \mathbb{C}$ with maximal and minimal imaginary part, i.e., the y-coordinate. We think of the y-coordinate as defining a height function on the curve. By applying Rolle's theorem to the height function $y(s)$, we obtain a minimum $s_{\min}$ and a maximum $s_{\max}$. The points $\alpha(s_{\min})$ and $\alpha(s_{\max})$ have "opposite" horizontal tangent vectors. Such points correspond to the values $\theta = 0$ and $\theta = \pi$. We have $G(\alpha(s_{\min})) = e^{i0}, \quad G(\alpha(s_{\max})) = e^{i\pi}$. This exhibits the points $e^{i0} \in S^1$ and $e^{i\pi} \in S^1$ as images of the Gauss map. To treat the general case, the idea is to use an analog of the height function which is, up to sign, the distance to the line $\mathrm{Span}_{\mathbb{R}} v$ spanned by the vector $v = e^{i\theta}$. Let us show how one obtains the pair of opposite tangent vectors

$$v = e^{i\theta} \;\text{ and }\; -v = e^{i(\theta+\pi)}. \tag{12.4.1}$$

Consider the vector

$$n_\theta = e^{i(\theta+\frac{\pi}{2})} \tag{12.4.2}$$

normal to v. Consider the function $h(s)$ (h for "height") given by the scalar product

$$h(s) = \langle \alpha(s), n_\theta \rangle \tag{12.4.3}$$

The function h is analogous to the y-coordinate in Step 2 above. We seek the extrema of the function h of (12.4.3). At a critical point s_0 of the function h, we have by Fermat's theorem, $\frac{d}{ds}\big|_{s=s_0} h(s) = \left\langle \frac{d\alpha}{ds}, n_\theta \right\rangle = 0$. where n_θ is the normal vector of (12.4.2). Hence the tangent vector $v(s_0) = \alpha'(s_0)$ at each critical point s_0 of h is parallel to the vector v of (12.4.1). The minimum and the maximum of h give the tangents $v = e^{i\theta}$ and $-v = e^{i(\theta+\pi)}$, which are therefore both in the image of the map G. As v ranges over S^1, we thus obtain the points on the curve where the tangent vector is parallel to v, provng surjectivity.

[Proof of the "one-to-one" part] We would like to show that the Gauss map is one-to-one. As before, we will work with tangent vectors in place of the normal vectors. We argue by contradiction. Suppose on the contrary that two distinct points $p \in J$ and $q \in J$ have identical tangent vectors $v(s) = e^{i\theta}$. Then the tangent lines T_p and T_q to J at p and q are parallel. By definition of convexity, the curve J lies entirely on one side of each of the tangent lines T_p and T_q. Since the lines are parallel, there are two possibilities: (1) the tangent lines coincide: $T_p = T_q$. (2) The curve lies in the strip between the two tangent lines. However, in the second case the tangent vectors at p and q have opposite directions v abd $-v$. It remains to treat the case $T_p = T_q$. Thus we can assume that both p and q must lie on the common line $T_p = T_q$. Therefore we obtain a straight line segment $[p, q] \subseteq J$. This contradicts the hypothesis that the curve J is *strictly* convex. The contradiction proves that the map is one-to-one.

12.5. Total curvature of a convex Jordan curve

The result on the total curvature[5] of a Jordan curve is of interest in its own right. Furthermore, the result serves to motivate an analogous statement of the Gauss–Bonnet theorem for surfaces in Section 11.6.

DEFINITION 12.5.1. The *total curvature* $\mathrm{Tot}(C)$ of a curve C with arclength parametrisation $\alpha(s)\colon [a,b] \to \mathbb{R}^2$ is the integral

$$\mathrm{Tot}(C) = \int_a^b k_\alpha(s)ds. \tag{12.5.1}$$

LEMMA 12.5.2. *The total curvature of a circle C_r (of radius r) is 2π and therefore indepedent of r.*

PROOF. Consider an arclength parametrisation

$$\alpha(s) : [0, 2\pi r] \to \mathbb{R}^2$$

of the circle C_r given by the usual trigonometric functions. Then

$$\mathrm{Tot}(C_r) = \int_0^{2\pi r} k_\alpha(s)ds = \int_0^{2\pi r} \frac{1}{r}\, ds = \frac{2\pi r}{r} = 2\pi.$$

□

DEFINITION 12.5.3 (Contour integral). If C is a twice differentiable smooth closed curve parametrized by $\alpha : [a,b] \to \mathbb{R}^2$, so that in particular $\alpha(a) = \alpha(b)$ and $\alpha'(a) = \alpha'(b)$, we will write the integral (12.5.1) using the notation of a contour integral

$$\mathrm{Tot}(C) = \oint_C k_\alpha(s)ds. \tag{12.5.2}$$

THEOREM 12.5.4. *Let C be a strictly convex smooth Jordan curve with arclength parametrisation $\alpha(s)$. Then the total curvature of C is $\mathrm{Tot}(C) = \oint_C k_\alpha(s)ds = 2\pi$.*

PROOF. Let $\alpha(s)$ be a unit speed parametrisation chosen so that $\alpha(0)$ is the lowest point (i.e., y is minimal) of the curve, and assume the curve is parametrized counterclockwise. Let

$$v(s) = \alpha'(s) \in \mathbb{C}.$$

Then $v(0) = e^{i0} = 1$ and therefore $\theta(0) = 0$. The function $\theta = \theta(s)$ is monotone increasing from 0 to 2π.[6] Applying the change of variable formula for integration, we obtain

$$\mathrm{Tot}(C) = \oint_C k_\alpha(s)ds = \oint_C \left|\frac{dv}{ds}\right| ds = \oint_C \frac{d\theta}{ds} ds = \int_0^{2\pi} d\theta = 2\pi,$$

[5] Akmumiyut kolelet

[6] By Theorem in Section 12.4.

proving the theorem. □

See also note.[7]

12.6. Total curvature of conic sections

Applying Theorem 12.5.4 we obtain the following result about conic sections. For the nondegeneracy condition of conic sections see Definition 2.6.5 and the main text there.

COROLLARY 12.6.1. *Consider a closed convex curve given by an ellipse $E \subseteq \mathbb{R}^2$ defined by a quadratic equation*

$$ax^2 + 2bxy + cy^2 + dx + ey + f = 0, \quad ac - b^2 > 0,$$

and assume that the ellipse is nondegenerate. Then

$$\mathrm{Tot}(E) = 2\pi.$$

Namely 2π is the total curvature of the ellipse E.

EXAMPLE 12.6.2 (Theorem inapplicable to hyperbola). The hyperbola H defined by $\lambda_1 x^2 + \lambda_2 y^2 = k$ is not a closed curve. In the special case $\lambda_1 = -\lambda_2$ (see Definition 2.6.2) when the asymptotes of H are orthogonal, the image of θ is precisely half the circle. Therefore we can define the total curvature of this non-closed curve by a similar integral, and a similar integration argument shows that the total curvature is π (rather than 2π).

REMARK 12.6.3 (Non-convex curves). A theorem similar to Theorem 12.5.4 (on total curvature) in fact holds for an arbitrary regular Jordan curve (even though in general $\theta(s)$ will not be a monotone function) *provided* we use *signed* curvature. In this section, we dealt only with the convex case in order to simplify the topological considerations. See further in Section 12.7.

REMARK 12.6.4. A similar calculation yields the Gauss–Bonnet theorem for convex surfaces in Section 11.6.

---

[7] We can also consider the normal vector $n(s)$ to the curve. The normal vector satisfies $n(s) = e^{i(\theta + \frac{\pi}{2})} = ie^{i(\theta)}$ and $|\frac{dn}{ds}| = |\frac{dv}{ds}| = \frac{d\theta}{ds}$. Therefore we can also calculate the total curvature as follows: $\oint_C k_\alpha(s) ds = \oint_C \left|\frac{dn}{ds}\right| ds = \oint_C \frac{d\theta}{ds} ds = \int_0^{2\pi} d\theta = 2\pi$, with $n(s)$ in place of $v(s)$. For surfaces, an analogous Gauss map is defined by means of the normal vector; see Section 11.4.

12.7. Rotation index of a closed curve in the plane

In this section we will deal with arbitrary closed curves (not necessarily convex). We start with the following data.

(1) $\alpha(s)$ is an arclength parametrisation of a closed curve in the plane.
(2) L is the length of the curve.
(3) The curve is parametrized counterclockwise.[8]
(4) $v(s) = \alpha'(s)$.

As in Section 12.4, we have the following result. The result applies to closed curves that are not necessarily simple (i.e., non-self-intersecting).

Theorem 12.7.1 (Existence of continuous branch of θ). *$\alpha(s)$ admits a continuous single-valued branch of the angle $\theta(s)$, $s \in [0, L]$, where $\theta(s)$ is the angle measured counterclockwise from the positive x-axis to the tangent vector $v(s) = \alpha'(s)$.*

If the closed curve is not convex, the function $\theta(s)$ will not be monotone and at certain points its derivative $\theta'(s)$ may be negative.

Once we have chosen a continuous branch of θ, we can define the signed curvature $\tilde{k}_\alpha$ of the parametrized closed curve as in Definition 12.2.1 by setting $\tilde{k}_\alpha(s) = \frac{d\theta}{ds}$ where $\theta = \frac{1}{i}\log v$, or equivalently $v(s) = \alpha'(s) = e^{i\theta(s)}$.

Lemma 12.7.2. *If C is a closed curve then the difference $\theta(L)-\theta(0)$ is necessarily an integer multiple of 2π.*

Proof. Consider a regular closed plane curve of length L with an arclength parametrisation $\alpha(s)$ with $\alpha'(s) = e^{i\theta(s)}$. By Theorem 12.7.1, a continuous branch of $\theta(s)$ can be chosen even if the curve is not simple. Such a branch is a map $\theta\colon [0, L] \to \mathbb{R}$. The values $\theta(L)$ and $\theta(0)$ must agree up to a multiple of 2π since they define the same tangent vector at the point $\alpha(0) = \alpha(L)$. □

Definition 12.7.3. The *rotation index* ι_α of a closed unit speed plane curve $\alpha(s)$ is the integer $\iota_\alpha = \frac{\theta(L)-\theta(0)}{2\pi}$.

Theorem 12.7.4. *The rotation index $\iota(J)$ of a smooth Jordan curve $J \subseteq \mathbb{C}$ is $\iota_J = \pm 1$ (namely, 1 for the counterclockwise orientation and -1 for the clockwise orientation).*

For a proof see Millman & Parker [**MP77**, p. 55]. We will analyze the rotation index further in Section 13.1.

[8] In the sense of the course on complex functions.

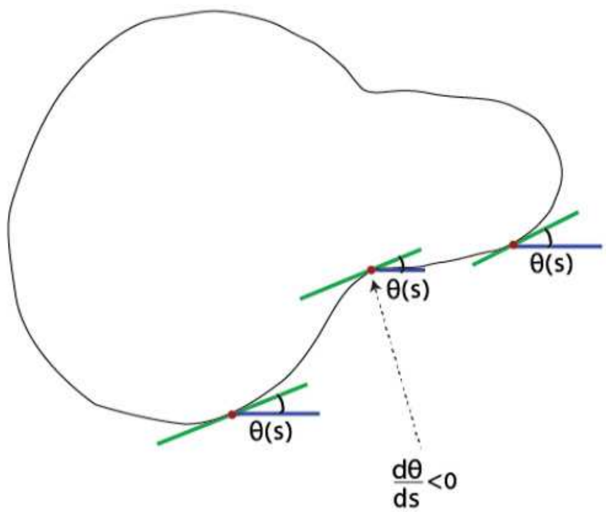

FIGURE 12.7.1. Three points on a curve with the same tangent direction, but the middle one corresponds to $\theta'(s) < 0$.

12.8. Total signed curvature of Jordan curve

We have the following generalisation of Theorem 12.5.4 on the total curvature.

DEFINITION 12.8.1. The *total signed curvature* of a smooth curve $C \subseteq \mathbb{R}^2$ of length L with arclength parametrisation $\alpha(s)$ for $s \in [0, L]$ is defined to be

$$\widetilde{\mathrm{Tot}}(C) = \int_0^L \tilde{k}_\alpha(s) ds.$$

We obtain the following generalisation of Theorem 12.5.4. As usual we assume that Jordan curves are parametrized counterclockwise.

THEOREM 12.8.2. *The total signed curvature of a smooth Jordan curve C oriented counterclockwise is*

$$\widetilde{\mathrm{Tot}}(C) = 2\pi. \tag{12.8.1}$$

PROOF. We exploit a continuous branch $\theta(s)$ as in Theorem 12.5.4, but *without* the absolute value signs on the derivative of $\theta(s)$. Applying the change of variable formula for integration, we obtain

$$\widetilde{\mathrm{Tot}}(C) = \oint_C \tilde{k}_\alpha(s)ds = \oint_C \frac{d\theta}{ds}ds = \int_{\theta(0)}^{\theta(0)+2\pi} d\theta = 2\pi,$$

where the upper limit of integration is $\theta(0)+2\pi$ by Theorem 12.7.4 on the rotation index. $\square$

See also note.[9]

12.9. Connected components of curves

Until now we have only considered connected curves. A curve may in general have several connected components (rechivei k'shirut), defined in Infi 3.[10] The curve C decomposes as a disjoint union

$$C = \bigsqcup_i C_i$$

of connected curves C_i. The set of connected components of C is denoted $\pi_0(C)$. The number of connected components is denoted $|\pi_0(C)|$.

The total curvature can be similarly defined for a non-connected curve, by summing the integrals over each connected components.

Definition 12.9.1. The total signed curvature of $C = \bigsqcup_i C_i$ is $\widetilde{\mathrm{Tot}}(C) = \sum_i \widetilde{\mathrm{Tot}}(C_i)$.

We obtain the following corollary of the theorem on total curvature of a Jordan curve, by applying the previous theorem to each connected component and summing the resulting total curvatures.

---

[9]There is a generalisation (to plane domains) of the formula (12.8.1) for the total signed curvature of a Jordan curve. The *Euler characteristic* $\chi(D)$ of a surface D, defined via a triangulation, is $\chi(D) = V - E + F$ (vertices minus edges plus faces). Theorem: Consider a plane domain $D \subseteq \mathbb{R}^2$ with (possibly several) smooth bundary components. Assume that each of the components of ∂D is oriented in a way compatible with the standard orientation in D, and let $\tilde{k}$ be the signed curvature. Then $\int_{\partial D} \tilde{k} = 2\pi\chi(D)$. We mention some examples. (1) If D is a disk then $\chi(D) = 1$ and total curvature of the boundary of D is 2π. (2) If D is an annulus then $\chi(D) = 0$ and total curvature of the boundary of D is 0. (3) If $D \subseteq \mathbb{R}^2$ has two holes (in addition to the outside boundary) then then $\chi(D) = -1$ and total curvature of the boundary is -2π. A more general Gauss–Bonnet theorem for surfaces is discussed in Sections 11.7 and 11.10. The compatibility of orientations is discussed in more detail in 88-826.

[10]Two points p, q on a curve $C \subseteq \mathbb{R}^2$ are said to lie in the same connected component if there exists a continuous map $h\colon [0,1] \to C$ such that $h(0) = p$ and $h(1) = q$. This defines an equivalence relation on the curve C.

Corollary 12.9.2. *If each connected component of a curve C is a Jordan curve parametrized counterclockwise, then the total signed curvature of C is* $\widetilde{\mathrm{Tot}}(C) = 2\pi\,|\pi_0(C)|$.

Example 12.9.3. Let $F(x,y) = (x^2+4y^2-1)((x-10)^2+4y^2-1)$, and let C_F be the curve defined by $F(x,y) = 0$. Then C_F is the union of a pair of disjoint ellipses. Therefore it has two connected components: $|\pi_0(C_F)| = 2$. By Corollary 12.6.1, each component has total curvature 2π. By Corollary 12.9.2, its total signed curvature is 4π.

12.10. Gaussian curvature as product of signed curvatures

We will exploit signed curvature defined in Section 12.1 to express Gaussian curvature as a product of signed curvatures of planar curves.

We originally defined the Gaussian curvature $K = K(u^1,u^2)$ as the determinant of the Weingarten map W_p at $p = x(u^1,u^2) \in M$.[11]

One can represent the Gaussian curvature as the product of signed curvatures of plane curves obtained by intersecting M with the planes spanned by n and each of the eigenvectors v_i, as already mentioned following Definition 9.5.1, in note (1). Then for signed curvatures one has the formula $K = \tilde{k}_{\beta_1}\tilde{k}_{\beta_2}$ as in Theorem 12.10.4 below. In this spirit, some textbooks (following Gauss himself) *define* the Gaussian curvature not as the determinant of the Weingarten map but rather as the product of the *signed* curvatures $\tilde{k}_\alpha$ of such a pair of plane curves. We consider the following data.

(1) $M \subseteq \mathbb{R}^3$ is a surface.
(2) $p \in M$.
(3) v_1, v_2 are orthonormal eigenvectors of the map $W_p\colon T_p \to T_p$.
(4) We choose a unit normal n to the surface M at $p \in M$.
(5) (v_1, v_2, n) is an orthonormal basis.

Definition 12.10.1 (Planes E_i). The plane $E_i = \mathrm{Span}(n, v_i) \subseteq \mathbb{R}^3$ is spanned by n and eigenvector v_i, for each $i = 1,2$.

Note that $E_i \cap T_p = \mathbb{R}v_i$, for $i = 1,2$.

Definition 12.10.2 (Plane curves). The plane curve $\beta_i = M \cap E_i$ is the intersection of the surface and the plane. Choose unit speed parametrisations $\beta_i(s)$ such that $\beta_i'(0) = v_i$.

[11]In note 7, we saw that the absolute value $|K|$ can be represented as the product of the curvatures k_{β_1} and k_{β_2} of geodesics in the direction of the eigenvectors v_i of W_p, namely $|K| = k_{\beta_1}k_{\beta_2}$. In this section, we sharpen this result so as to express the Gaussian curvature itself as a product of signed curvatures.

DEFINITION 12.10.3. The signed curvature $\tilde{k}_i$ (relative to n) of the curves $\beta_i(s)$ is $\tilde{k}_i = \langle \beta_i''(0), n \rangle$, for each $i = 1, 2$.

The signed curvatures are the eigenvalues of W_p (see Section 12.11 for more details). We therefore obtain the following theorem.

THEOREM 12.10.4. *The Gaussian curvature K_p of $M \subseteq \mathbb{R}^3$ at p satisfies*

$$K_p = \tilde{k}_1 \tilde{k}_2 \tag{12.10.1}$$

where $\tilde{k}_i$, $i = 1, 2$ is the signed curvature of the plane curves β_1 and β_2.

12.11. Connection to the Hessian

To explain formula (12.10.1) in terms of the Hessian, we can assume without loss of generality that M is the graph of a function $f(x, y)$ vanishing at the origin, with a critical point at the origin. Furthermore, we can assume without loss of generality that the eigendirections of the Hessian of f coincide with the x-axis and the y-axis (by orthogonally diagonalizing the quadratic part of f as in Theorem 2.5.3).[12]

PROPOSITION 12.11.1. *The Gaussian curvature at the origin of the graph of f is the product of signed curvatures of the curves obtained as the graphs of the restrictions of f respectively to the x-axis and the y-axis.*

PROOF. By hypothesis, the Taylor formula with remainder for f gives $f(x, y) = ax^2 + by^2 + o(x^2 + y^2)$. In the notation of Definition 12.10.2, the curve $\beta_1 \subseteq E_1$ in the (x, z) plane is the graph of $z = ax^2 + o(x^2)$. The curve $\beta_2 \subseteq E_2$ in the (y, z) plane is the graph of $z = by^2 + o(y^2)$. With respect to the normal $n = e_3$, the curves have *signed* curvature respectively $2a$ and $2b$ at the origin. The Hessian is the diagonal martix $H_f = \begin{pmatrix} 2a & 0 \\ 0 & 2b \end{pmatrix}$. Therefore $K = 4ab$ as required. □

REMARK 12.11.2. The value $2a$ is simultaneously the partial derivative f_{xx}, the signed curvature of the curve β_1, and also the eigenvalue of the Hessian and of the Weingarten map corresponding to the eigenvector e_1 (similarly, e_2 belongs to the eigenvalue $2b$).

	Weingarten map $(L^i{}_j)$	Gaussian curvature K	mean curvature H
plane	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$	0	0
cylinder	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	0	$\frac{1}{2}$
invariance of curvature		yes	no

TABLE 12.12.1. Plane and cylinder have the same intrinsic geometry (K), but different extrinsic geometries (H)

12.12. Mean versus Gaussian curvature

Recall that by Gauss' *theorema egregium* the Gaussian curvature can be expressed in terms of the metric coefficients alone.

THEOREM 12.12.1. *Unlike Gaussian curvature K, the mean curvature $H = \frac{1}{2}L^i{}_i$ cannot be expressed in terms of the metric coefficients g_{ij} and their derivatives.*

Namely, the plane and the cylinder have parametrisations with identical (g_{ij}), but with different mean curvature, *cf.* Table 12.12.1. To summarize, Gaussian curvature is an intrinsic invariant, while mean curvature an extrinsic invariant, of the surface.[13]

[12] In more detail, if the eigenspaces are not the axes, we adjust the coordinates by a suitable rotation of $\mathbb{R}^3$ so that the x and y axes become the eigenspaces, while the z-axis is the direction of the normal. Note that Gaussian curvature is unchanged under such a rotation. as under all orthogonal transformations.

[13] [Algebraic degree] This material is optional. Let $\alpha(s)$ be a parametrisation of a smooth closed curve C. Consider an arbitrary smooth map (not necessarily the Gauss map) $\alpha(s) \mapsto e^{i\theta(s)}$ from C to the circle. The *algebraic degree* of the map at a point $z \in S^1$ is defined to be the sum $\sum_{\theta^{-1}(z)} \operatorname{sign}\left(\frac{d\theta}{ds}\right)$, where the summation is over all points in the inverse image of z. Here by Sard's theorem z can be chosen in such a way that the inverse image is finite so that the sum is well-defined. Theorem: For Jordan curves (i.e., embedded curves), the algebraic degree of the Gauss map is 1 if oriented counterclockwise and -1 if orientated counterclockwise. For nonconvex Jordan curves, the Gauss map to the circle will not be one-to-one, but will still have an algebraic degree one. This phenomenon has an analogue for embedded surfaces in $\mathbb{R}^3$ where the algebraic degree is proportional to the Euler characteristic. Example: If the curve is not simple, the degree may be different from ± 1. Thus, the map defined by $z \mapsto z^n$ restricted to the unit circle gives a map $e^{i\theta} \mapsto e^{in\theta}$ of algebraic degree n. The preimage of $1 = e^{i0}$ consists precisely of

the nth roots of unity $e^{\frac{2\pi i k}{n}}, k = 0, 1, 2, \ldots, n-1$. Altogether there are n of them and therefore the algebraic degree is n.

CHAPTER 13

Rotation index, isothermisation, pseudosphere

13.1. Rotation index of closed curves and total curvature

The rotation index was defined in Section 12.7 where we mostly dealt with Jordan curves. For general regular closed curves (with possible self-intersections) we have the following result.

The result is analogous to the relation between the Euler characteristic of a surface and its total Gaussian curvature (see Section 11.10).

THEOREM 13.1.1. *Let $\alpha(s)$ be an arclength parametrisation of a geometric curve $C \subseteq \mathbb{R}^2$. Then the rotation index ι_α is related to the total signed curvature $\widetilde{\mathrm{Tot}}(C)$ as follows:*

$$\widetilde{\mathrm{Tot}}(C) = 2\pi\iota_\alpha. \tag{13.1.1}$$

PROOF. The proof is similar to that given in Section 12.7. We exploit the continuous branch $\theta(s) = \frac{1}{i}\log\alpha'(s)$, $s \in [0, L]$ as in Theorem 12.5.4. Applying the change of variable formula for integration, we obtain

$$\widetilde{\mathrm{Tot}}(C) = \oint_C \tilde{k}_\alpha(s)ds = \oint_C \frac{d\theta}{ds}ds = \int_0^{2\pi\iota_\alpha} d\theta = 2\pi\iota_\alpha,$$

where the upper limit of integration is $2\pi\iota_\alpha$ (by definition of the rotation index). $\square$

EXAMPLE 13.1.2. The rotation index of a figure-8 curve (known as the lemniscate of Gerono, one of the Lissajous curves) is 0. Hence its total signed curvature vanishes.

EXAMPLE 13.1.3. Describe an immersed curve with rotation index 2.

13.2. Surfaces of revolution and isothermalisation

Below we will express the metric of a surface of revolution in isothermal coordinates. Recall that if a surface is obtained by revolving a curve $(r(\phi), z(\phi))$, we obtain metric coefficients $g_{11} = r^2$ and $g_{22} = \left(\frac{dr}{d\phi}\right)^2 + \left(\frac{dz}{d\phi}\right)^2$. Thus we obtain the following theorem.

THEOREM 13.2.1. *We have the following expression of the metric of a surface of revolution in coordinates $u^1 = r$, $u^2 = \phi$:*

$$g_{11} = r^2, \quad g_{22} = \left(\tfrac{dr}{d\phi}\right)^2 + \left(\tfrac{dz}{d\phi}\right)^2, \quad g_{12} = 0. \tag{13.2.1}$$

We will use this theorem to carry out explicit isothermalisation in the case of a surface of revolution.

PROPOSITION 13.2.2. *Let $(r(\phi), z(\phi))$, where $r(\phi) > 0$, be an arclength parametrisation of the generating curve of a surface of revolution M. Then the change of variable*

$$\psi = \int \frac{1}{r(\phi)} d\phi,$$

produces an isothermal *parametrisation of M in terms of variables (θ, ψ). With respect to the new coordinates, the first fundamental form is given by a scalar matrix with metric coefficients $r(\phi(\psi))^2 \delta_{ij}$.*

REMARK 13.2.3. The existence of such a parametrisation is predicted by the uniformisation theorem (Theorem 13.12.2) in the case of a general surface but the general result is less explicit.

PROOF OF PROPOSITION 13.2.2. We will carry out a change of variables affecting only the second variable ϕ.

Step 1. Consider an arbitrary monotone change of parameter $\phi = \phi(\psi)$. By chain rule,

$$\frac{dr}{d\psi} = \frac{dr}{d\phi}\frac{d\phi}{d\psi}.$$

Step 2. Consider the (possibly non-scalar) first fundamental form as given by (13.2.1). We need to impose the condition

$$g_{11} = g_{22} \tag{13.2.2}$$

to ensure that the matrix of metric coefficients be a scalar matrix. By Theorem 13.2.1, equality (13.2.2) is equivalent to the equation

$$r^2 = \left(\frac{dr}{d\psi}\right)^2 + \left(\frac{dz}{d\psi}\right)^2. \tag{13.2.3}$$

Step 3. By chain rule, (13.2.3) is equivalent to the formula $r^2 = \left(\left(\tfrac{dr}{d\phi}\right)^2 + \left(\tfrac{dz}{d\phi}\right)^2\right)\left(\tfrac{d\phi}{d\psi}\right)^2$, or

$$r = \sqrt{\left(\frac{dr}{d\phi}\right)^2 + \left(\frac{dz}{d\phi}\right)^2}\,\frac{d\phi}{d\psi}. \tag{13.2.4}$$

Step 4. In the case when the original generating curve is parametrized by arclength (as in our proposition), equation (13.2.4) becomes $r = \frac{d\phi}{d\psi}$. Solving for ψ, we obtain $\psi(\phi) = \int \frac{d\phi}{r(\phi)}$.

Step 5. Since the change of parameter is monotone, we can solve the equation $\psi = \psi(\phi)$ for ϕ obtaining $\phi = \phi(\psi)$. Substituting the new variable $\phi(\psi)$ in place of ϕ in the parametrisation of the surface, we obtain a new parametrisation of the surface of revolution in coordinates (θ, ψ).

Step 6. With respect to the new parametrisation, the first fundamental form is scalar with conformal factor $\lambda = r^2(\phi(\psi))$ as per equation (13.2.3). □

13.3. Gaussian Curvature of pseudosphere

As an application of the isothermalisation of Section 13.2, we calculate the curvature of the pseudosphere. The pseudosphere (see Section 5.9) is the surface of revolution generated by functions $r(\phi) = e^\phi$ and $z(\phi) = \int_0^\phi \sqrt{1 - e^{2\tau}} d\tau = -\int_\phi^0 \sqrt{1 - e^{2\tau}} d\tau$, where $-\infty < \phi \leq 0$. Its metric coefficients are given by the matrix $(g_{ij}) = \begin{pmatrix} e^{2\phi} & 0 \\ 0 & 1 \end{pmatrix}$, which is not a scalar matrix.

Theorem 13.3.1. *The pseudosphere has constant Gaussian curvature $K = -1$.*

Proof. Instead of applying the general formula for curvature, we use the trick of a change of coordinates that results in isothermal coordinates so we can apply the formula for curvature in terms of the Laplace–Beltrami operator. We apply Proposition 13.2.2 to introduce the change of coordinates

$$\psi = \int \frac{d\phi}{r(\phi)} = \int e^{-\phi} d\phi = -e^{-\phi}. \tag{13.3.1}$$

From (13.3.1) we obtain $\phi = -\ln(-\psi)$. This results in isothermal coordinates (θ, ψ) with conformal factor $f(\psi) = r(\phi(\psi))$. Therefore

$$f(\psi) = r(\phi(\psi)) = e^{-\ln(-\psi)} = \frac{1}{e^{\ln(-\psi)}} = -\frac{1}{\psi}.$$

We have $\lambda = f^2 = \frac{1}{\psi^2}$. Recall that the metric with $g_{ij}(\theta, \psi) = \frac{1}{\psi^2}\delta_{ij}$ is by definition hyperbolic (here we merely use the variables θ, ψ in place of x, y). Therefore it has curvature $K = -1$ as shown in Section 10.2. For completeness we carry out the calculation as follows. In isothermal coordinates (θ, ψ), Gaussian curvature is given by the formula in terms of the Laplace–Beltrami operator. The Gaussian curvature of the pseudosphere is

$$\begin{aligned} K &= -\Delta_{LB} \ln f \\ &= -\frac{1}{\lambda}\frac{\partial^2}{\partial \psi^2} \ln f(\psi) \\ &= -\psi^2 \frac{\partial^2}{\partial \psi^2} \ln((-\psi)^{-1}) \\ &= \psi^2 \frac{\partial^2}{\partial \psi^2} \ln(-\psi). \end{aligned}$$

Differentiating with respect to ψ, we obtain

$$\begin{aligned} K &= \psi^2 \frac{\partial}{\partial \psi}\left(\frac{-1}{-\psi}\right) \\ &= \psi^2 \frac{\partial}{\partial \psi}\left(\frac{1}{\psi}\right) \\ &= \psi^2 \left(-\frac{1}{\psi^2}\right) \\ &= -1, \end{aligned}$$

as required. □

13.4. Transition from classical to modern diff geom

REMARK 13.4.1. The theorema egregium of Gauss marks the transition from classical differential geometry of curves and surfaces embedded in 3-space, to modern differential geometry of surfaces (and manifolds) studied intrinsically.

More specifically, once we have a notion of Gaussian curvature (and more generally sectional curvature) that only depends on the metric on a surface (or manifold), we can study the geometry of the surface (or manifold) *intrinsically*, i.e., without any reference to a Euclidean embedding.

REMARK 13.4.2. To formulate the intrinsic viewpoint, one needs the notion of duality of vector and covector. This theme is treated in Section 13.5.

13.5. Duality in linear algebra; 1-forms

Let V be a finite-dimensional real vector space.

EXAMPLE 13.5.1. Euclidean space $\mathbb{R}^n$ is an example of a real vector space of dimension n with basis $(e_1, \ldots, e_n)$.

EXAMPLE 13.5.2. The tangent plane T_pM of a regular surface M at a point $p \in M$ is a real vector space of dimension 2.

DEFINITION 13.5.3. A *linear form*, also called *1-form*, ϕ on V is a linear functional

$$\phi : V \to \mathbb{R}.$$

DEFINITION 13.5.4 (dx, dy). In the usual Euclidean plane of vectors

$$v = v^1 e_1 + v^2 e_2$$

represented by arrows, we denote by dx the 1-form which extracts the abscissa of the vector, and by dy the 1-form which extracts the ordinate of the vector:

$$dx(v) = v^1,$$

and

$$dy(v) = v^2.$$

EXAMPLE 13.5.5. For a vector $v = 3e_1 + 4e_2$ with components $(3, 4)$ we obtain

$$dx(v) = 3, \ dy(v) = 4.$$

DEFINITION 13.5.6. The quadratic forms dx^2 and dy^2 are defined by squaring the value of the 1-form on v:

$$dx^2(v) = (dx(v))^2.$$

Such quadratic forms are called *rank-1* quadratic forms.

Thus,

$$dx^2(v) = (v^1)^2, \ dy^2(v) = (v^2)^2.$$

EXAMPLE 13.5.7. In the case $v = 3e_1 + 4e_2$ we obtain $dx^2(v) = 9$ and $dy^2(v) = 16$.

13.6. Dual vector space

DEFINITION 13.6.1. The *dual* space of V, denoted V^*, is the space of all linear forms on V:

$$V^* = \{\phi \colon \phi \text{ is a 1-form on } V\}.$$

Evaluating ϕ at an element $x \in V$ produces a scalar $\phi(x) \in \mathbb{R}$.

DEFINITION 13.6.2. The *evaluation map*[1] is the natural pairing between V and V^*, namely a linear map denoted

$$\langle \ , \ \rangle : V \times V^* \to \mathbb{R},$$

defined by evaluating y at x, i.e., setting $\langle x, y \rangle = y(x)$, for all $x \in V$ and $y \in V^*$.

[1] hatzava?

REMARK 13.6.3. Note we are using the same notation for the pairing as for the scalar product in Euclidean space. The notation is quite widespread.

DEFINITION 13.6.4. If V admits a basis of vectors

$$(x_i) = (x_1, x_2, \ldots, x_n),$$

then the dual space V^* admits a unique basis, called the *dual basis* $(y_j) = (y_1, \ldots, y_n)$, satisfying

$$\langle x_i, y_j \rangle = \delta_{ij}, \tag{13.6.1}$$

for all $i, j = 1, \ldots, n$, where δ_{ij} is the Kronecker delta function.

EXAMPLE 13.6.5. Let $V = \mathbb{R}^2$. We have the standard basis e_1, e_2 for V. The 1-forms dx, dy form a basis for the dual space V^*. Then the basis (dx, dy) is the dual basis to the basis (e_1, e_2).

EXAMPLE 13.6.6. Let $V = \mathbb{R}^2$ identified with $\mathbb{C}$ for convenience. We have the basis $(1, e^{\frac{i\pi}{3}})$ for V where $1 = e_1 + 0e_2$ while

$$e^{\frac{i\pi}{3}} = \frac{1}{2}e_1 + \frac{\sqrt{3}}{2}e_2.$$

Find the dual basis in V^*. The answer is $y_1 = dx - \frac{1}{\sqrt{3}}dy$, $y_2 = \frac{2}{\sqrt{3}}dy$.

13.7. Derivations

Let E be a space of dimension n, and let $p \in E$ be a fixed point. The following definition is independent of coordinates.

DEFINITION 13.7.1. Let

$$\mathbb{D}_p = \{f \colon f \in C^\infty\}$$

be the ring of real C^∞-functions f defined in an arbitrarily small open neighborhood of $p \in E$.

REMARK 13.7.2. The ring operations are pointwise multiplication and pointwise addition, where we choose the intersection of the two domains as the domain of the new function.

Note that $\mathbb{D}_p$ is infinite-dimensional as it includes all polynomials.

THEOREM 13.7.3. *Choose coordinates* $(u^1, \ldots, u^n)$ *in* E. *A partial derivative* $\frac{\partial}{\partial u^i}$ *at* p *is a* 1*-form*

$$\frac{\partial}{\partial u^i} \colon \mathbb{D}_p \to \mathbb{R}$$

on the space $\mathbb{D}_p$, satisfying Leibniz rule

$$\left.\frac{\partial(fg)}{\partial u^i}\right|_p = \left.\frac{\partial f}{\partial u^i}\right|_p g(p) + f(p)\left.\frac{\partial g}{\partial u^i}\right|_p \tag{13.7.1}$$

for all $f, g \in \mathbb{D}_p$.

Formula (13.7.1) can be written briefly as

$$\tfrac{\partial}{\partial u^i}(fg) = \tfrac{\partial}{\partial u^i}(f)g + f\tfrac{\partial}{\partial u^i}(g),$$

with the understanding that both sides are evaluated at the point p. This was proved in calculus. Formula (13.7.1) motivates the following more general definition of a derivation.

DEFINITION 13.7.4. A *derivation* X at the point $p \in E$ is a 1-form

$$X\colon \mathbb{D}_p \to \mathbb{R}$$

on the space $\mathbb{D}_p$ satisfying Leibniz rule:

$$X(fg) = X(f)g(p) + f(p)X(g) \tag{13.7.2}$$

for all $f, g \in \mathbb{D}_p$.

REMARK 13.7.5. Linearity of a derivation is required only with regard to scalars in $\mathbb{R}$, not with respect to functions.

13.8. Characterisation of derivations

It turns out that the space of derivations is spanned by partial derivatives.

PROPOSITION 13.8.1. *Let E be an n-dimensional space, and $p \in E$. Then the collection of all derivations at p is a vector space of dimension n.*

PROOF IN CASE $n = 1$. We will first prove the result in the case $n = 1$ of a single variable u at the point $p = 0$. Let $X\colon \mathbb{D}_p \to \mathbb{R}$ be a derivation. Then $X(1) = X(1 \cdot 1) = 2X(1)$ by Leibniz rule. Therefore $X(1) = 0$, and similarly for any constant by linearity of X.

Now consider the monic polynomial $u = u^1$ of degree 1, viewed as a linear function $u \in \mathbb{D}_{p=0}$. We evaluate the derivation X at the element $u \in \mathbb{D}$ and set $c = X(u)$. By the Taylor remainder formula, any function $f \in \mathbb{D}_{p=0}$ can be written as $f(u) = a + bu + g(u)u$ where $b =$

$\frac{\partial f}{\partial u}(0)$ and g is smooth and $g(0)=0$. Now we have by linearity

$$\begin{aligned} X(f) &= X(a+bu+g(u)u) \\ &= bX(u) + X(g)u(0) + g(0)\cdot X(u) \\ &= bc + 0 + 0 \\ &= c\,\frac{\partial}{\partial u}(f). \end{aligned}$$

Thus the derivation X coincides with the derivation $c\frac{\partial}{\partial u}$ for all $f \in \mathbb{D}_p$. Hence the tangent space is 1-dimensional and spanned by the element $\frac{\partial}{\partial u}$, proving the theorem in this case. □

[2]

13.9. Tangent space and cotangent space

DEFINITION 13.9.1. Let $p \in E$. The space of derivations at p is called the *tangent space* $T_p = T_pE$ at p.

The results of the previous section can be formulated as follows.

---

[2]**Proof in case $n = 2$.** This material is optional. Let us prove the result in the case $n=2$ of two variables u, v at the origin $p=(0,0)$. Let X be a derivation. Then $X(1) = X(1\cdot 1) = 2X(1)$ by Leibniz rule. Therefore $X(1)=0$, and similarly for any constant by linearity of X. Now consider the monic polynomial $u = u^1$ of degree 1, i.e., the linear function $u \in \mathbb{D}_{p=0}$. We evaluate the derivation X at u and set $c = X(u)$. Similarly, consider the monic polynomial $v = v^1$ of degree 1, i.e., the linear function $v \in \mathbb{D}_{p=0}$. We evaluate the derivation X at v and set $\tilde{c} = X(v)$. By the Taylor remainder formula, any function $f \in \mathbb{D}_{p=0}$, where $f = f(u,v)$, can be written as $f(u,v) = a + bu + \tilde{b}v + g(u,v)u^2 + h(u,v)uv + k(u,v)v^2$, where the functions $g(u,v)$, $h(u,v)$, and $k(u,v)$ are smooth. Note that the coefficients b and $\tilde{b}$ are the first partial derivatives of f at the origin. Now we have

$$\begin{aligned} X(f) &= X(a+bu+\tilde{b}v+g(u,v)u^2+h(u,v)uv+k(u,v)v^2) \\ &= bX(u) + \tilde{b}X(v) + \big(X(g)u^2 + g(u,v)2uX(u) + \cdots\big) \\ &= bX(u) + \tilde{b}X(v) \\ &= bc + \tilde{b}\tilde{c} \\ &= c\frac{\partial}{\partial u}(f) + \tilde{c}\frac{\partial}{\partial v}(f) \end{aligned}$$

by evaluating at the point $(0,0)$. Thus the derivation X coincides with the derivation $c\frac{\partial}{\partial u}+\tilde{c}\frac{\partial}{\partial v}$ for all test functions $f \in \mathbb{D}_p$. Hence the two partials span the tangent space. Therefore the tangent space is 2-dimensional, proving the theorem in this case.

COROLLARY 13.9.2. *Let $(u^1, \ldots, u^n)$ be coordinates for E, and let $p \in E$. Then a basis for the tangent space T_p is given by the n partial derivatives*

$$\left(\frac{\partial}{\partial u^i}\right), \quad i = 1, \ldots, n.$$

DEFINITION 13.9.3. The space dual to the tangent space T_p is called the *cotangent space*, and denoted T_p^*.

DEFINITION 13.9.4. An element of a tangent space is a vector, while an element of a cotangent space is called a 1-form, or a *covector*.

DEFINITION 13.9.5. The basis dual to the basis $\left(\frac{\partial}{\partial u^i}\right)$ is denoted

$$(du^j), \quad j = 1, \ldots, n.$$

Thus each du^j is by definition a linear form, or cotangent vector (covector for short). We are therefore working with dual bases $\left(\frac{\partial}{\partial u^i}\right)$ for vectors, and (du^j) for covectors. The evaluation map as in (13.6.1) gives

$$\left\langle \frac{\partial}{\partial u^i}, du^j \right\rangle = du^j \left(\frac{\partial}{\partial u^i}\right) = \delta_i^j, \tag{13.9.1}$$

where δ_i^j is the Kronecker delta.

13.10. Constructing bilinear forms out of 1-forms

Recall that the polarisation formula (see Definition 1.4.3) allows one to reconstruct a symmetric bilinear form $B = B(v, w)$, from the quadratic form $Q(v) = B(v, v)$, at least if the characteristic is not 2:

$$B(v, w) = \frac{1}{4}(Q(v + w) - Q(v - w)). \tag{13.10.1}$$

Similarly, one can construct bilinear forms out of the 1-forms du^i, as follows.

EXAMPLE 13.10.1. Consider a quadratic form $a_i(du^i)^2$ defined by a linear combination of the rank-1 quadratic forms $(du^i)^2$, as in Definition 13.5.6.

Polarizing the quadratic form, one obtains a bilinear form on the tangent space T_p.

EXAMPLE 13.10.2. Let $\underline{v} = v^1 e_1 + v^2 e_2$ be an arbitrary vector in the plane. Let dx and dy be the standard covectors, extracting, respectively, the first and second coordinates of v. Consider the quadratic form Q given by $Q = Edx^2 + Fdy^2$, where $E, F \in \mathbb{R}$. Here $Q(\underline{v})$ is calculated as

$$Q(\underline{v}) = E(dx(\underline{v}))^2 + F(dy(\underline{v}))^2 = E(v^1)^2 + F(v^2)^2.$$

Polarisation then produces the bilinear form $B = B(\underline{v}, \underline{w})$, where $\underline{v}$ and $\underline{w}$ are arbitrary vectors, given by the formula

$$B(\underline{v}, \underline{w}) = E\, dx(\underline{v})\, dx(\underline{w}) + F\, dy(\underline{v})\, dy(\underline{w}).$$

EXAMPLE 13.10.3. Setting $E = F = 1$ in the previous example, we obtain the standard scalar product in the plane:

$$B(v, w) = v \cdot w = dx(v)\, dx(w) + \, dy(v)\, dy(w) = v^1 w^1 + v^2 w^2.$$

13.11. First fundamental form

DEFINITION 13.11.1. A metric (or first fundamental form) g is a symmetric bilinear form on the tangent space at p, namely $g \colon T_p \times T_p \to \mathbb{R}$, defined for all p and varying continuously and smoothly in p.

REMARK 13.11.2. In Riemannian geometry one requires the associated quadratic form to be positive definite. In relativity theory one uses a form of type $(3, 1)$.

Recall that the basis for T_p in coordinates (u^i) is given by the tangent vectors $\frac{\partial}{\partial u^i}$. These are given by certain derivations (see Section 13.7).

The first fundamental form g is traditionally expressed by a matrix of coefficients called *metric coefficients* g_{ij}, giving the inner product of the i-th and the j-th vector in the basis: $g_{ij} = g\left(\frac{\partial}{\partial u^i}, \frac{\partial}{\partial u^j}\right)$, where g is the first fundamental form.

DEFINITION 13.11.3. The square norm the i-th vector is given by the coefficient $g_{ii} = \left\| \frac{\partial}{\partial u^i} \right\|^2$.

We will express the first fundamental form in more intrinsic notation of quadratic forms built from 1-forms (covectors).

13.12. Dual bases in differential geometry

Let us now restrict attention to the case of 2 dimensions, i.e., the case of surfaces. At every point $p = (u^1, u^2)$, we have the metric coefficients $g_{ij} = g_{ij}(u^1, u^2)$. Each metric coefficient is thus a function of two variables.

We will only consider the case when the matrix is diagonal. This can always be achieved, in two dimensions, at a point by a suitable change of coordinates, by the uniformisation theorem (see Theorem 13.12.2).

We set $x = u^1$ and $y = u^2$ to simplify notation. In the notation developed in Section 13.10, we can write the quadratic form associated with the first fundamental form as follows:

$$g = g_{11}(x,y)(dx)^2 + g_{22}(x,y)(dy)^2. \tag{13.12.1}$$

For example, if the metric coefficients form an identity matrix: $g_{ij} = \delta_{ij}$, we obtain the standard flat metric

$$g = (dx)^2 + (dy)^2 \tag{13.12.2}$$

or simply $g = dx^2 + dy^2$.

Example 13.12.1 (Hyperbolic metric). Let $g_{11} = g_{22} = \frac{1}{y^2}$ at each point (x, y) where $y > 0$. This means that $\left|\frac{\partial}{\partial u^1}\right| = \frac{1}{y}$ and $\left|\frac{\partial}{\partial u^2}\right| = \frac{1}{y}$. The resulting hyperbolic metric in the upper half plane $\{y > 0\}$ is expressed by the quadratic form $\frac{1}{y^2}(dx^2 + dy^2)$. Note that this expression is undefined whenever $y = 0$. See Section 10.2. The hyperbolic metric in the upper half plane is a complete metric.

Closely related results are the Riemann mapping theorem and the conformal representation theorem.

Theorem 13.12.2 (Riemann mapping/uniformisation). *Every metric on a connected*[3] *surface is conformally equivalent to a metric of constant Gaussian curvature.*

From the complex analytic viewpoint, the uniformisation theorem states that every Riemann surface is covered by either the sphere, the plane, or the upper halfplane. Thus no notion of curvature is needed for the statement of the uniformisation theorem. However, from the differential geometric point of view, what is relevant is that every conformal class of metrics contains a metric of constant Gaussian curvature. See [**Ab81**] for a lively account of the history of the uniformisation theorem. A lot of information on the uniformisation theorem and the Riemann mapping theorem can be found at `https://mathoverflow.net/q/10516`.

13.13. More on dual bases

Recall that if $(x_1, \ldots, x_n)$ is a basis for a vector space V then the dual vector space V^* possesses a basis called the *dual basis* and denoted $(y_1, \ldots, y_n)$ satisfying $\langle x_i, y_j \rangle = y_j(x_i) = \delta_{ij}$.

[3]kashir

EXAMPLE 13.13.1. In $\mathbb{R}^2$ we have a basis $(x_1, x_2) = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$ in the tangent plane T_p at a point p. The dual basis of 1-forms (y_1, y_2) for T_p^* is denoted (dx, dy). Thus we have

$$\left\langle \frac{\partial}{\partial x}, dx \right\rangle = dx\left(\frac{\partial}{\partial x}\right) = 1$$

and

$$\left\langle \frac{\partial}{\partial y}, dy \right\rangle = dy\left(\frac{\partial}{\partial y}\right) = 1,$$

while

$$\left\langle \frac{\partial}{\partial x}, dy \right\rangle = dy\left(\frac{\partial}{\partial x}\right) = 0,$$

etcetera.

Similarly, in polar coordinates at a point $p \neq 0$ we have a basis $\left(\frac{\partial}{\partial r}, \frac{\partial}{\partial \theta}\right)$ for T_p, and a dual basis $(dr, d\theta)$ for T_p^*. Thus we have

$$\left\langle \frac{\partial}{\partial r}, dr \right\rangle = dr\left(\frac{\partial}{\partial r}\right) = 1$$

and

$$\left\langle \frac{\partial}{\partial \theta}, d\theta \right\rangle = d\theta\left(\frac{\partial}{\partial \theta}\right) = 1,$$

while

$$\left\langle \frac{\partial}{\partial r}, d\theta \right\rangle = d\theta\left(\frac{\partial}{\partial r}\right) = 0,$$

etcetera.

Now in polar coordinates we have a natural area element $r\,dr\,d\theta$. Area of a region D is calculated by Fubini's theorem as

$$\iint_D r\,dr\,d\theta = \int\left(\int r dr\right) d\theta.$$

Thus we have a natural basis $(y_1, y_2) = (rdr, d\theta)$ in T_p^* when $p \neq 0$, i.e., $y_1 = rdr$ while $y_2 = d\theta$. Its dual basis (x_1, x_2) in T_p can be easily identified. It is

$$(x_1, x_2) = \left(\frac{1}{r}\frac{\partial}{\partial r}, \frac{\partial}{\partial \theta}\right).$$

Indeed, we have

$$\langle x_1, y_1 \rangle = \left\langle \frac{1}{r}\frac{\partial}{\partial r}, rdr \right\rangle = rdr\left(\frac{1}{r}\frac{\partial}{\partial r}\right) = r\frac{1}{r}\,dr\left(\frac{\partial}{\partial r}\right) = 1,$$

etcetera.

CHAPTER 14

Lattices and tori

14.1. Circle via the exponential map

We will discuss lattices in Euclidean space $\mathbb{R}^b$ in Section 14.2. Here we give an intuitive introduction in the simplest case $b = 1$. Every lattice (discrete[1] subgroup; see definition below in Section 14.2) in $\mathbb{R} = \mathbb{R}^1$ is of the form

$$L_\alpha = \alpha\mathbb{Z} = \{n\alpha : n \in \mathbb{Z}\} \subseteq \mathbb{R},$$

for some real $\alpha > 0$. It is spanned by the vector αe_1 (or $-\alpha e_1$).

The lattice $L_\alpha \subseteq \mathbb{R}$ is in fact an additive subgroup. Therefore we can form the quotient group $\mathbb{R}/L_\alpha$. This quotient is a circle (see Theorem 14.1.1). The 1-volume, *i.e.* the length, of the quotient circle is precisely α. We will give a description in terms of the complex function e^z.

THEOREM 14.1.1. *The quotient group $\mathbb{R}/L_\alpha$ is isomorphic to the circle $S^1 \subseteq \mathbb{C}$.*

PROOF. Consider the map $\hat{\phi} : \mathbb{R} \to \mathbb{C}$ defined by

$$\hat{\phi}(x) = e^{\frac{i2\pi x}{\alpha}}.$$

By the usual addition rule for the exponential function, this map is a homomorphism from the additive structure on $\mathbb{R}$ to the multiplicative structure in the group $\mathbb{C} \setminus \{0\}$. Namely, we have

$$(\forall x \in \mathbb{R})(\forall y \in \mathbb{R})\ \hat{\phi}(x + y) = \hat{\phi}(x)\hat{\phi}(y).$$

Furthermore, we have $\hat{\phi}(x + \alpha m) = \hat{\phi}(x)$ for all $m \in \mathbb{Z}$ and therefore $\ker \hat{\phi} = L_\alpha$. By the group-theoretic isomorphism theorem, the map $\hat{\phi}$ descends to a map

$$\phi : \mathbb{R}/L_\alpha \to \mathbb{C},$$

which is injective. Its image is the unit circle $S^1 \subseteq \mathbb{C}$, which is a group under multiplication. □

[1] b'didah

14.2. Lattice, fundamental domain

Let $b > 0$ be an integer.

Definition 14.2.1. A *lattice* $L \subseteq \mathbb{R}^b$ is the integer span of a linearly independent set of b vectors.

Thus, if vectors $v_1, \ldots, v_b$ are linearly independent, then they span a lattice

$$L = \{n_1 v_1 + \ldots + n_b v_b : n_i \in \mathbb{Z}\} = \mathbb{Z}v_1 + \mathbb{Z}v_2 \cdots + \mathbb{Z}v_b$$

Note that the subgroup is isomorphic to $\mathbb{Z}^b$.

Definition 14.2.2. An *orbit* of a point $x_0 \in \mathbb{R}^b$ under the action of a lattice L is the subset of $\mathbb{R}^b$ given by the collection of elements

$$\{x_0 + g : g \in L\}.$$

These can also be viewed as the cosets of the lattice in $\mathbb{R}^b$.

Definition 14.2.3. The quotient

$$\mathbb{R}^b/L$$

is called a b-torus.

At this point tori are understood at the group-theoretic level as in the case of the circle $\mathbb{R}/L$.

14.3. Fundamental domain

Definition 14.3.1. A *fundamental domain* for the torus $\mathbb{R}^b/L$ is a closed set $F \subseteq \mathbb{R}^b$ satisfying the following three conditions:

- every orbit meets F in at least one point;
- every orbit meets the interior $\mathrm{Int}(F)$ of F in at most one point;
- the boundary ∂F is of zero b-dimensional volume (and can be thought of as a union of $(b-1)$-dimensional hyperplanes).

In the literature, one often replaces "n-dimensional volume" by n-dimensional "Lebesgue measure".

Example 14.3.2. The interval $[0, \alpha]$ is a fundamental domain for the circle $\mathbb{R}/L_\alpha$.

Example 14.3.3. The parallelepiped spanned by a collection of basis vectors for $L \subseteq \mathbb{R}^b$ is a fundamental domain fot L.

More concretely, consider the following example.

EXAMPLE 14.3.4. The vectors e_1 and e_2 in $\mathbb{R}^2$ span the unit square which is a fundamental domain for the lattice $\mathbb{Z}^2 \subseteq \mathbb{R}^2 = \mathbb{C}$ of Gaussian integers.

EXAMPLE 14.3.5. Consider the vectors $v = (1, 0)$ and $w = (\frac{1}{2}, \frac{\sqrt{3}}{2})$ in $\mathbb{R}^2 = \mathbb{C}$. Their span is a parallelogram giving a fundamental domain for the lattice of Eisenstein integers (see Example 14.4.2).

DEFINITION 14.3.6. The *total volume* of the b-torus $\mathbb{R}^b/L$ is by definition the b-volume of a fundamental domain.

It is shown in advanced calculus that the total volume thus defined is independent of the choice of a fundamental domain.

14.4. Lattices in the plane

Let $b = 2$. Every lattice $L \subseteq \mathbb{R}^2$ is of the form

$$L = \mathrm{Span}_{\mathbb{Z}}(v, w) \subseteq \mathbb{R}^2,$$

where $\{v, w\}$ is a linearly independent set. For example, let α and β be nonzero reals. Set

$$L_{\alpha,\beta} = \mathrm{Span}_{\mathbb{Z}}(\alpha e_1, \beta e_2) \subseteq \mathbb{R}^2.$$

This lattice admits an orthogonal basis, namely $\{\alpha e_1, \beta e_2\}$.

EXAMPLE 14.4.1 (Gaussian integers). For the standard lattice $\mathbb{Z}^b \subseteq \mathbb{R}^b$, the torus $\mathbb{T}^b = \mathbb{R}^b/\mathbb{Z}^b$ satisfies $\mathrm{vol}(\mathbb{T}^b) = 1$ as it has the unit cube as a fundamental domain.

In dimension 2, the resulting lattice in $\mathbb{C} = \mathbb{R}^2$ is called the *Gaussian integers* L_G. It contains 4 elements of least length. These are the fourth roots of unity. We have

$$L_G = \mathrm{Span}_Z(1, i) \subseteq \mathbb{C}.$$

EXAMPLE 14.4.2 (Eisenstein integers). Consider the lattice $L_E \subseteq \mathbb{R}^2 = \mathbb{C}$ spanned by $1 \in \mathbb{C}$ and the sixth root of unity $e^{\frac{2\pi i}{6}} \in \mathbb{C}$:

$$L_E = \mathrm{Span}_{\mathbb{Z}}(e^{i\pi/3}, 1) = \mathbb{Z}\, e^{i\pi/3} + \mathbb{Z}\, 1 \subseteq \mathbb{C}. \tag{14.4.1}$$

The resulting lattice is called the *Eisenstein integers*. The torus $\mathbb{T}^2 = \mathbb{R}^2/L_E$ satisfies $\mathrm{area}(\mathbb{T}^2) = \frac{\sqrt{3}}{2}$. The Eisenstein lattice contains 6 elements of least length, namely all the sixth roots of unity.

14.5. Successive minima of a lattice

Let B be Euclidean space, and let $\| \ \|$ be the Euclidean norm. Let $L \subseteq (B, \| \ \|)$ be a lattice, i.e., span of a collection of b linearly independent vectors where $b = \dim(B)$.

DEFINITION 14.5.1. The first successive minimum, $\lambda_1(L, \| \ \|)$ is the least length of a nonzero vector in L.

We can express the definition symbolically by means of the formula

$$\lambda_1(L, \| \ \|) = \min \left\{ \|v_1\| \ \middle| \ v_1 \in L \setminus \{0\} \right\}.$$

We illustrate the geometric meaning of λ_1 in terms of the circle of Theorem 14.1.1.

THEOREM 14.5.2. *Consider a lattice $L \subseteq \mathbb{R}$. Then the circle $\mathbb{R}/L$ satisfies*

$$\mathrm{length}(\mathbb{R}/L) = \lambda_1(L).$$

PROOF. This follows by choosing the fundamental domain $F = [0, \alpha]$ where $\alpha = \lambda_1(L)$, so that $L = \alpha\mathbb{Z}$, cf. Example 14.2 above. $\square$

REMARK 14.5.3. When $\alpha = 1$, we can choose a representative from the orbit of x to be the fractional part $\{x\}$ of x.

DEFINITION 14.5.4. For $k = 2$, define the *second successive minimum* of the lattice L with $rank(L) \geq 2$ as follows. Given a pair of vectors $S = \{v, w\}$ in L, define the *size*[2] $|S|$ of S by setting

$$|S| = \max(\|v\|, \|w\|).$$

Then the second successive minimum, $\lambda_2(L, \| \ \|)$ is the least size of a pair of non-proportional vectors in L:

$$\lambda_2(L) = \inf_S |S|,$$

where S runs over all linearly independent (*i.e.* non-proportional) pairs of vectors $\{v, w\} \subseteq L$.

EXAMPLE 14.5.5. For both the Gaussian and the Eisenstein integers we have $\lambda_1 = \lambda_2 = 1$.

EXAMPLE 14.5.6. For the lattice $L_{\alpha,\beta}$ we have $\lambda_1(L_{\alpha,\beta}) = \min(|\alpha|, |\beta|)$ and $\lambda_2(L_{\alpha,\beta}) = \max(|\alpha|, |\beta|)$.

[2]Quotation marks: merka'ot.

14.6. Gram matrix

The volume of the torus $\mathbb{R}^b/L$ (see Definition 14.2.3) is also called the covolume of the lattice L. It is by definition the volume of a fundamental domain for L, *e.g.* a parallelepiped spanned by a $\mathbb{Z}$-basis for L.

The Gram matrix was defined in Definition 5.5.3.

THEOREM 14.6.1. *Let $L \subset \mathbb{R}^b$ be a lattice spanned by linearly independent vectors $(v_1, \ldots, v_b)$. Then the volume of the torus $\mathbb{R}^b/L$ is the square root of the determinant of the Gram matrix $Gram(v_1, \ldots, v_b)$.*

In geometric terms, the parallelepiped P spanned by the vectors $\{v_i\}$ satisfies

$$\operatorname{vol}(P) = \sqrt{\det(Gram(S))}. \tag{14.6.1}$$

PROOF. Let A be the square matrix whose columns are the column vectors $v_1, v_2 \ldots, v_n$ in $\mathbb{R}^n$. It is shown in linear algebra that

$$\operatorname{vol}(P) = |\det(A)|.$$

Let $B = A^t A$, and let $B = (b_{ij})$. Then

$$b_{ij} = v_i^t\, v_j = \langle v_i, v_j \rangle$$

Hence $B = Gram(S)$. Thus

$$\det(Gram(S)) = \det(A^t\ A) = \det(A)^2 = \operatorname{vol}(P)^2$$

proving the theorem. □

14.7. Sphere and torus as topological surfaces

The topology of surfaces will be discussed in more detail in Chapter 17. For now, we will recall that a compact surface can be either orientable or non-orientable. An orientable surface is characterized topologically by its *genus*, i.e. number of "handles".

Recall that the unit sphere in $\mathbb{R}^3$ can be represented implicitly by the equation

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

Parametric representations of surfaces are discussed in Section 5.2.

EXAMPLE 14.7.1. The sphere has genus 0 (no handles).

THEOREM 14.7.2. *The 2-torus is characterized topologically in one of the following four equivalent ways:*

(1) *the Cartesian product of a pair of circles: $S^1 \times S^1$;*
(2) *the surface of revolution in $\mathbb{R}^3$ obtained by starting with the following circle in the (x, z)-plane: $(x - 10)^2 + z^2 = 1$ (for example), and rotating it around the z-axis;*

(3) *a quotient* $\mathbb{R}^2/L$ *of the plane by a lattice* L*;*
(4) *a compact* 2*-dimensional manifold of genus* 1*.*

The equivalence between items (2) and (3) can be seen by marking a pair of generators of L by different color, and using the same colors to indicate the corresponding circles on the embedded torus of revolution, as follows:

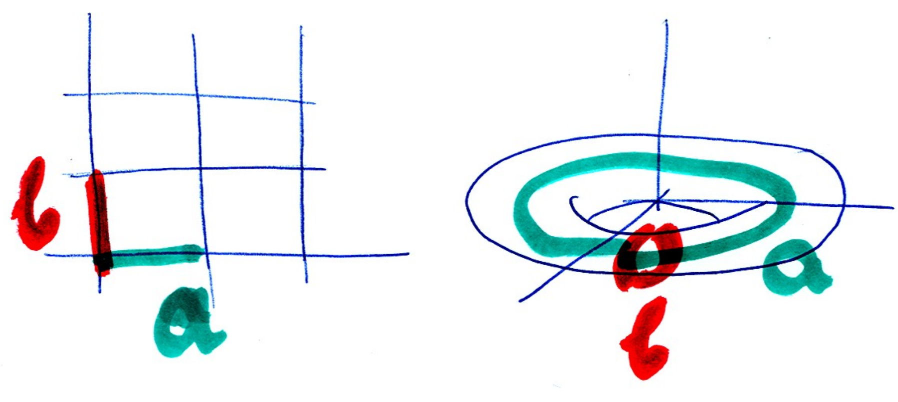

FIGURE 14.7.1. Torus viewed by means of its lattice (left) and by means of a Euclidean embedding (right)

Note by comparison that a circle can be represented either by its fundamental domain which is $[0, 2\pi]$ (with endpoints identified), or as the unit circle embedded in the plane.[3]

14.8. Standard fundamental domain

We will discuss the case $b = 2$ in detail. An important role is played in this dimension by the standard fundamental domain.

---

[3]The Hermite constant γ_b is defined in one of the following two equivalent ways:

(1) γ_b is the *square* of the maximal first successive minimum λ_1, among all lattices of unit covolume;
(2) γ_b is defined by the formula

$$\sqrt{\gamma_b} = \sup \left\{ \frac{\lambda_1(L)}{\operatorname{vol}(\mathbb{R}^b/L)^{\frac{1}{b}}} \,\middle|\, L \subseteq (\mathbb{R}^b, \|\ \|) \right\}, \tag{14.7.1}$$

where the supremum is extended over all lattices L in $\mathbb{R}^b$ with a Euclidean norm $\|\ \|$.

A lattice realizing the supremum may be thought of as the one realizing the densest packing in $\mathbb{R}^b$ when we place the balls of radius $\frac{1}{2}\lambda_1(L)$ at the points of L. In dimensions $b \geq 3$, the Hermite constants are harder to compute, but explicit values (as well as the associated critical lattices) are known for small dimensions, *e.g.* $\gamma_3 = 2^{\frac{1}{3}} = 1.2599...$, while $\gamma_4 = \sqrt{2} = 1.4142....$

DEFINITION 14.8.1. The *standard fundamental domain*, denoted D, is the set

$$D = \left\{ z \in \mathbb{C} \;\middle|\; |z| \geq 1,\ |\mathrm{Re}(z)| \leq \frac{1}{2},\ \mathrm{Im}(z) > 0 \right\} \tag{14.8.1}$$

cf. [**Ser73**, p. 78].

The domain D a fundamental domain for the action of $PSL(2, \mathbb{Z})$ in the upper half-plane of $\mathbb{C}$.

LEMMA 14.8.2. *Multiplying a lattice $L \subseteq \mathbb{C}$ by nonzero complex numbers does not change the value of the quotient*

$$\frac{\lambda_1(L)^2}{\mathrm{area}(\mathbb{C}/L)}.$$

PROOF. We write such a complex number as $re^{i\theta}$. Note that multiplication by $re^{i\theta}$ can be thought of as a composition of a scaling by the real factor r, and rotation by angle θ. The rotation is an isometry (congruence) that preserves all lengths, and in particular the length $\lambda_1(L)$ and the area of the quotient torus.

Meanwhile, multiplication by r results in a cancellation

$$\frac{\lambda_1(rL)^2}{\mathrm{area}(\mathbb{C}/rL)} = \frac{(r\lambda_1(L))^2}{r^2\,\mathrm{area}(\mathbb{C}/L)} = \frac{r^2\lambda_1(L)^2}{r^2\,\mathrm{area}(\mathbb{C}/L)} = \frac{\lambda_1(L)^2}{\mathrm{area}(\mathbb{C}/L)},$$

proving the lemma. □

14.9. Conformal parameter τ of a lattice

Two lattices in $\mathbb{C}$ are said to be *similar* if one is obtained from the other by multiplication by a nonzero complex number.

THEOREM 14.9.1. *Every lattice in $\mathbb{C}$ is similar to a lattice spanned by $\{\tau, 1\}$ where τ is in the standard fundamental domain D of* (14.8.1). *The value $\tau = e^{i\pi/3}$ corresponds to the Eisenstein integers* (14.4.1).

PROOF. Let $L \subseteq \mathbb{C}$ be a lattice. Choose a "shortest" vector $z \in L$, *i.e.* we have $|z| = \lambda_1(L)$. By Lemma 14.8.2, we may replace the lattice L by the lattice $z^{-1}L$.

Thus, we may assume without loss of generality that the complex number $+1 \in \mathbb{C}$ is a shortest element in the lattice L. Thus we have $\lambda_1(L) = 1$. Now complete the element $+1$ to a $\mathbb{Z}$-basis

$$\{\bar{\tau}, +1\}$$

for L. Here we may assume, by replacing $\bar{\tau}$ by $-\bar{\tau}$ if necessary, that $\mathrm{Im}(\bar{\tau}) > 0$.

Now consider the real part $\mathrm{Re}(\bar{\tau})$. We adjust the basis by adding a suitable integer k to $\bar{\tau}$:

$$\tau = \bar{\tau} - k \quad \text{where} \quad k = \left[\mathrm{Re}(\bar{\tau}) + \tfrac{1}{2}\right] \tag{14.9.1}$$

(the brackets denote the integer part), so it satisfies the condition

$$-\frac{1}{2} \leq \mathrm{Re}(\tau) \leq \frac{1}{2}.$$

Since $\tau \in L$, we have $|\tau| \geq \lambda_1(L) = 1$. Therefore the element τ lies in the standard fundamental domain (14.8.1). □

EXAMPLE 14.9.2. For the "rectangular" lattice $L_{\alpha,\beta} = \mathrm{Span}_{\mathbb{Z}}(\alpha, \beta i)$, we obtain

$$\tau(L_{\alpha,\beta}) = \begin{cases} \frac{|\beta|}{|\alpha|} i & \text{if } |\beta| > |\alpha| \\ \frac{|\alpha|}{|\beta|} i & \text{if } |\alpha| > |\beta|. \end{cases}$$

COROLLARY 14.9.3. *Let $b = 2$. Then we have the following value for the Hermite constant: $\gamma_2 = \frac{2}{\sqrt{3}} = 1.1547....$ The corresponding optimal lattice is homothetic to the $\mathbb{Z}$-span of cube roots of unity in $\mathbb{C}$ (i.e. the Eisenstein integers).*

PROOF. Choose τ as in (14.9.1) above. The pair

$$\{\tau, +1\}$$

is a basis for the lattice. The imaginary part satisfies $\mathrm{Im}(\tau) \geq \frac{\sqrt{3}}{2}$, with equality possible precisely for

$$\tau = e^{i\frac{\pi}{3}} \text{ or } \tau = e^{i\frac{2\pi}{3}}.$$

Moreover, if $\tau = r\exp(i\theta)$, then

$$|\tau| \sin\theta = \mathrm{Im}(\tau) \geq \frac{\sqrt{3}}{2}.$$

The proof is concluded by calculating the area of the parallelogram in $\mathbb{C}$ spanned by τ and $+1$;

$$\frac{\lambda_1(L)^2}{\mathrm{area}(\mathbb{C}/L)} = \frac{1}{|\tau|\sin\theta} \leq \frac{2}{\sqrt{3}},$$

proving the theorem. □

DEFINITION 14.9.4. A $\tau \in D$ is said to be the *conformal parameter* of a flat torus T^2 if T^2 is similar to a torus $\mathbb{C}/L$ where $L = \mathbb{Z}\tau + \mathbb{Z}1$.

14.10. Conformal parameter τ of tori of revolution

The results of Section 13.2 have the following immediate consequence.

COROLLARY 14.10.1. *Consider a torus of revolution in $\mathbb{R}^3$ formed by rotating a Jordan curve of length $L > 0$, with unit speed parametisation $(f(\phi), g(\phi))$ where $\phi \in [0, L]$. Then the torus is conformally equivalent to a flat torus*

$$\mathbb{R}^2/L_{c,d}.$$

Here $\mathbb{R}^2$ is the (θ, ψ)-plane, where ψ is the antiderivative of $\frac{1}{f(\phi)}$ as in Section 13.2; while the rectangular lattice $L_{c,d} \subset \mathbb{R}^2$ is spanned by the orthogonal vectors $c\frac{\partial}{\partial\theta}$ and $d\frac{\partial}{\partial\psi}$, so that

$$L_{c,d} = \text{Span}\left(c\frac{\partial}{\partial\theta}\ ,\ d\frac{\partial}{\partial\psi}\right) = c\mathbb{Z} \oplus d\mathbb{Z},$$

where $c = 2\pi$ and $d = \int_0^L \frac{d\phi}{f(\phi)}$.

FIGURE 14.10.1. Torus: lattice (left) and embedding (right)

In Section 14.8 we showed that every flat torus $\mathbb{C}/L$ is similar to the torus spanned by $\tau \in \mathbb{C}$ and $1 \in \mathbb{C}$, where τ is in the standard fundamental domain

$$D = \{z = x + iy \in \mathbb{C} : |x| \le \tfrac{1}{2}, y > 0, |z| \ge 1\}.$$

DEFINITION 14.10.2. The parameter τ is called the *conformal parameter* of the torus.

COROLLARY 14.10.3. *The conformal parameter τ of a torus of revolution is pure imaginary:*

$$\tau = i\sigma^2$$

of absolute value

$$\sigma^2 = \max\left\{\frac{c}{d}\,,\ \frac{d}{c}\right\} \geq 1.$$

Proof. The proof is immediate from the fact that the lattice is rectangular. □

14.11. θ-loops and ϕ-loops on tori of revolution

Consider a torus of revolution (T^2, g) generated by a Jordan curve C in the (x,z)-plane, i.e., by a simple loop C, parametrized by a pair of functions $f(\phi)$, $g(\phi)$, so that $x = f(\phi)$ and $z = g(\phi)$.

Definition 14.11.1. A ϕ-loop on the torus is a simple loop obtained by fixing the coordinate θ (i.e., the variable ϕ is changing). A θ-loop on the torus is a simple loop obtained by fixing the coordinate ϕ (i.e., the variable θ is changing).

Proposition 14.11.2. *All ϕ-loops on the torus of revolution have the same length equal to the length L of the generating curve C (see Corollary 14.10.1).*

Proof. The surface is rotationally invariant. In other words, all rotations around the z-axis are isometries. Therefore all ϕ-loops have the same length. □

Proposition 14.11.3. *The θ-loops on the torus of revolution have variable length, depending on the ϕ-coordinate of the loop. Namely, the length is $2\pi x = 2\pi f(\phi)$.*

Proof. The proof is immediate from the fact that the function $f(\phi)$ gives the distance r to the z-axis. □

Definition 14.11.4. We denote by λ_ϕ the (common) length of all ϕ-loops on a torus of revolution.

Definition 14.11.5. We denote by $\lambda_{\theta_{\min}}$ the least length of a θ-loop on a torus of revolution, and by $\lambda_{\theta_{\max}}$ the maximal length of such a θ-loop.

14.12. Tori generated by round circles

Let $a, b > 0$. We assume $a > b$ so as to obtain tori that are embedded in 3-space. We consider the 2-parameter family $g_{a,b}$ of tori of revolution in 3-space with circular generating loop. The torus of revolution $g_{a,b}$ generated by a round circle is the locus of the equation

$$(r-a)^2 + z^2 = b^2, \tag{14.12.1}$$

where $r = \sqrt{x^2 + y^2}$. Note that the angle θ of the cylindrical coordinates (r, θ, z) does not appear in the equation (14.12.1). The torus is obtained by rotating the circle

$$(x-a)^2 + z^2 = b^2 \tag{14.12.2}$$

around the z-axis in $\mathbb{R}^3$. The torus admits a parametrisation in terms of the functions[4] $f(\phi) = a + b\cos\phi$ and $g(\phi) = b\sin\phi$. Namely, we have

$$x(\theta, \phi) = ((a + b\cos\phi)\cos\theta, (a + b\cos\phi)\sin\theta, b\sin\phi). \tag{14.12.3}$$

Here the θ-loop (see Section 14.11) has length $2\pi(a + b\cos\phi)$. The shortest θ-loop is therefore of length

$$\lambda_{\theta\,\min} = 2\pi(a-b),$$

and the longest one is

$$\lambda_{\theta_{\max}} = 2\pi(a+b).$$

Meanwhile, the ϕ-loop has length

$$\lambda_\phi = 2\pi b.$$

14.13. Conformal parameter of tori of revolution, residues

This section is optional.

We would like to compute the conformal parameter τ of the standard tori as in (14.12.3). We first modify the parametrisation so as to obtain a generating curve parametrized by arclength:

$$f(\varphi) = a + b\cos\frac{\varphi}{b}, \quad g(\varphi) = b\sin\frac{\varphi}{b}, \tag{14.13.1}$$

where $\varphi \in [0, L]$ with $L = 2\pi b$.

THEOREM 14.13.1. *The corresponding flat torus is given by the lattice L in the (θ, ψ) plane of the form $L = \operatorname{Span}_{\mathbb{Z}}\left(c\frac{\partial}{\partial\theta}\,,\, d\frac{\partial}{\partial\phi}\right)$ where $c = 2\pi$ and $d = \frac{2\pi}{\sqrt{(a/b)^2-1}}$. Thus the conformal parameter τ of the flat torus satisfies $\tau = i\max\left(((a/b)^2-1)^{-1/2}, ((a/b)^2-1)^{1/2}\right)$.*

PROOF. By Corollary 14.10.3, replacing φ by $\varphi(\psi)$ produces isothermal coordinates (θ, ψ) for the torus generated by (14.13.1), where $\psi = \int \frac{d\varphi}{f(\varphi)} = \int \frac{d\varphi}{a + b\cos\frac{\varphi}{b}}$, and therefore the flat metric is defined by a lattice in the (θ, ψ) plane with $c = 2\pi$ and $d = \int_0^{L=2\pi b} \frac{d\varphi}{a+b\cos\frac{\varphi}{b}}$. Changing the variable to to $\phi = \frac{\varphi}{b}$ we obtain $d = \int_0^{2\pi} \frac{d\phi}{(a/b)+\cos\phi}$, where $a/b > 1$.

[4]We use ϕ here and φ for the modified arclength parameter in the next section

Let $\alpha = a/b$. Now the integral is the real part Re of the complex integral $d = \int_0^{2\pi} \frac{d\phi}{\alpha+\cos\phi} = \int \frac{d\phi}{\alpha+\mathrm{Re}(e^{i\phi})}$. Thus

$$d = \int \frac{2d\phi}{2\alpha + e^{i\phi} + e^{-i\phi}}. \tag{14.13.2}$$

The change of variables $z = e^{i\phi}$ yields $d\phi = \frac{-idz}{z}$ and along the circle we have $d = \oint \frac{-2idz}{z(2\alpha+z+z^{-1})} = \oint \frac{-2idz}{z^2+2\alpha z+1} = \oint \frac{-2idz}{(z-\lambda_1)(z-\lambda_2)}$, where $\lambda_1 = -\alpha + \sqrt{\alpha^2-1}$ and $\lambda_2 = -\alpha - \sqrt{\alpha^2-1}$. The root λ_2 is outside the unit circle. Hence we need the residue at λ_1 to apply the residue theorem. The residue at λ_1 equals $\mathrm{Res}_{\lambda_1} = \frac{-2i}{\lambda_1-\lambda_2} = \frac{-2i}{2\sqrt{\alpha^2-1}} = \frac{-i}{\sqrt{\alpha^2-1}}$. The integral is determined by the residue theorem in terms of the residue at the pole $z = \lambda_1$. Therefore the lattice parameter d from Corollary 14.10.3 can be computed from (14.13.2) as $d = \left(2\pi i \,\mathrm{Res}_{\lambda_1}\right) = \frac{2\pi}{\sqrt{(\alpha)^2-1}}$. proving the theorem. $\square$

CHAPTER 15

A hyperreal view

15.1. Successive extensions $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, ${}^*\mathbb{R}$

Our reference for true infinitesimal calculus is Keisler's textbook [**Ke74**], downloadable at
`http://www.math.wisc.edu/~keisler/calc.html`

We start by motivating the familiar sequence of extensions of number systems

$$\mathbb{N} \hookrightarrow \mathbb{Z} \hookrightarrow \mathbb{Q} \hookrightarrow \mathbb{R}$$

in terms of their applications in arithmetic, algebra, and geometry. Each successive extension is introduced for the purpose of solving problems, rather than enlarging the number system for its own sake. Thus, the extension $\mathbb{Q} \subseteq \mathbb{R}$ enables one to express the length of the diagonal of the unit square and the area of the unit disc in our number system.

The familiar continuum $\mathbb{R}$ is an Archimedean continuum, in the sense that it satisfies the following Archimedean property.

Definition 15.1.1. An ordered field extending $\mathbb{N}$ is said to satisfy the *Archimedean property* if

$$(\forall \epsilon > 0)(\exists n \in \mathbb{N})\, [n\epsilon > 1]\,.$$

We will provisionally denote the real continuum $\mathbb{A}$ where "A" stands for *Archimedean*. Thus we obtain a chain of extensions

$$\mathbb{N} \hookrightarrow \mathbb{Z} \hookrightarrow \mathbb{Q} \hookrightarrow \mathbb{A},$$

as above. In each case one needs an enhanced ordered[1] number system to solve an ever broader range of problems from algebra or geometry.

The next stage is the extension

$$\mathbb{N} \hookrightarrow \mathbb{Z} \hookrightarrow \mathbb{Q} \hookrightarrow \mathbb{A} \hookrightarrow \mathbb{B},$$

where $\mathbb{B}$ is a *Bernoullian continuum* containing infinitesimals, defined as follows.

Definition 15.1.2. A Bernoullian extension of $\mathbb{R}$ is any *proper* extension which is an ordered field.

[1] sadur

Any Bernoullian extension allows us to define infinitesimals and do interesting things with those. But things become really interesting if we assume the Transfer Principle (Section 15.4), and work in a true hyperreal field, defined as in Definition 15.3.4 below. We will provide some motivating comments for the transfer principle in Section 15.3.

15.2. Motivating discussion for infinitesimals

Infinitesimals can be motivated from three different angles: geometric, algebraic, and arithmetic/analytic.

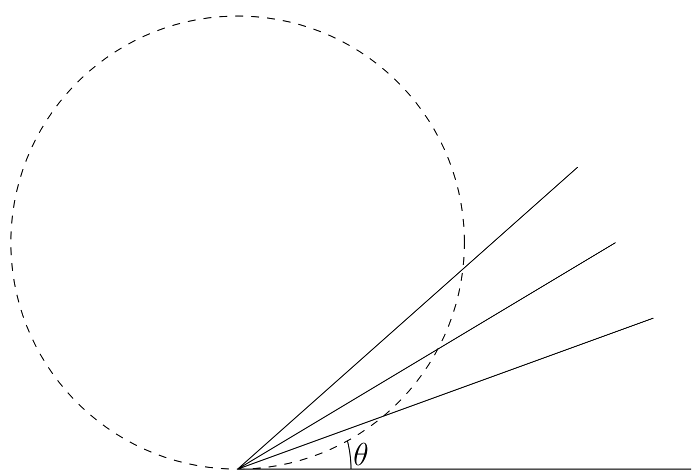

FIGURE 15.2.1. Horn angle θ is smaller than every rectilinear angle

(1) **Geometric (horn angles):** Some students have expressed the sentiment that they did not understand infinitesimals until they heard a geometric explanation of them in terms of what was classically known as horn angles. A horn angle is the crivice between a circle and its tangent line at the point of tangency. If one thinks of this crevice as a quantity, it is easy to convince oneself that it should be smaller than every rectilinear angle (see Figure 15.2.1). This is because a sufficiently small arc of the circle will be contained in the convex region cut out by the rectilinear angle no matter how small. When one renders this in terms of analysis and arithmetic, one gets a positive quantity smaller than every positive real number. We cite this example merely as intuitive motivation (our actual construction is different).

(2) **Algebraic (passage from ring to field):** The idea is to represent an infinitesimal by a sequence tending to zero. One can get something in this direction without reliance on any form of the axiom of choice. Namely, take the ring S of all sequences of real numbers, with arithmetic operations defined term-by-term. Now quotient the ring S by the equivalence relation that declares two sequences to be equivalent if they differ only on a finite set of indices. The resulting object S/K is a proper ring extension of $\mathbb{R}$, where $\mathbb{R}$ is embedded by means of the constant sequences. However, this object is not a field. For example, it has zero divisors. But quotienting it further in such a way as to get a field, by extending the kernel K to a *maximal* ideal K', produces a field S/K', namely a hyperreal field.

(3) **Analytic/arithmetic:** One can mimick the construction of the reals out of the rationals as the set of equivalence classes of Cauchy sequences, and construct the hyperreals as equivalence classes of sequences of real numbers under an appropriate equivalence relation.

15.3. Introduction to the transfer principle

The *transfer principle* is a type of theorem that, depending on the context, asserts that rules, laws or procedures valid for a certain number system, still apply (i.e., are "transfered") to an extended number system.

EXAMPLE 15.3.1. The familiar extension $\mathbb{Q} \subseteq \mathbb{R}$ preserves the property of being an ordered field.

EXAMPLE 15.3.2. To give a negative example, the extension $\mathbb{R} \subseteq \mathbb{R} \cup \{\pm\infty\}$ of the real numbers to the so-called *extended reals* does not preserve the property of being an ordered field.

The hyperreal extension $\mathbb{R} \subseteq {}^*\mathbb{R}$ (defined below) preserves *all* first-order properties (i.e., properties involving quantification over elements but not over sets.

EXAMPLE 15.3.3. The formula $\sin^2 x + \cos^2 x = 1$, true over $\mathbb{R}$ for all real x, remains valid over ${}^*\mathbb{R}$ for all hyperreal x, including infinitesimal and infinite values of $x \in {}^*\mathbb{R}$.

Thus the transfer principle for the extension $\mathbb{R} \subseteq {}^*\mathbb{R}$ is a theorem asserting that any statement true over $\mathbb{R}$ is similarly true over ${}^*\mathbb{R}$, and vice versa. Historically, the transfer principle has its roots in the procedures involving Leibniz's *Law of continuity.*[2]

We will explain the transfer principle in several stages of increasing degree of abstraction. More details can be found in Section 15.4.

DEFINITION 15.3.4. An ordered field $\mathbb{B}$, properly including the field $\mathbb{A} = \mathbb{R}$ of real numbers (so that $\mathbb{A} \subsetneqq \mathbb{B}$) and satisfying the Transfer Principle, is called a hyperreal field. If such an extended field $\mathbb{B}$ is fixed then elements of $\mathbb{B}$ are called hyperreal numbers,[3] while the extended field itself is usually denoted ${}^*\mathbb{R}$.

THEOREM 15.3.5. *Hyperreal fields exist.*

For example, a hyperreal field can be constructed as the quotient of the ring $\mathbb{R}^{\mathbb{N}}$ of sequences of real numbers, by an appropriate maximal ideal.

DEFINITION 15.3.6. A positive infinitesimal is a positive hyperreal number ϵ such that

$$(\forall n \in \mathbb{N})\,[n\epsilon < 1]$$

More generally, we have the following.

DEFINITION 15.3.7. A hyperreal number ε is said to be *infinitely small* or *infinitesimal* if

$$-a < \varepsilon < a$$

for every positive real number a.

In particular, one has $\varepsilon < \frac{1}{2}, \quad \varepsilon < \frac{1}{3}, \quad \varepsilon < \frac{1}{4}, \quad \varepsilon < \frac{1}{5}$, etc. If $\varepsilon > 0$ is infinitesimal then $N = \frac{1}{\varepsilon}$ is positive infinite, i.e., greater than every real number.

A hyperreal number that is not an infinite number are called *finite.* Sometimes the term *limited* is used in place of *finite.*

Keisler's textbook exploits the technique of representing the hyperreal line graphically by means of dots indicating the separation between the finite realm and the infinite realm. One can view infinitesimals with microscopes as in Figure 15.6.1. One can also view infinite numbers with telescopes as in Figure 15.3.1. We have an important subset

$$\{\text{finite hyperreals}\} \subseteq {}^*\mathbb{R}$$

---

[2]Leibniz's theoretical strategy in dealing with infinitesimals was analyzed in a number of detailed studies recently, which found Leibniz' strategy to be more robust than George Berkeley's flawed critique thereof.

[3]Similar terminology is used with regard to integers and hyperintegers.

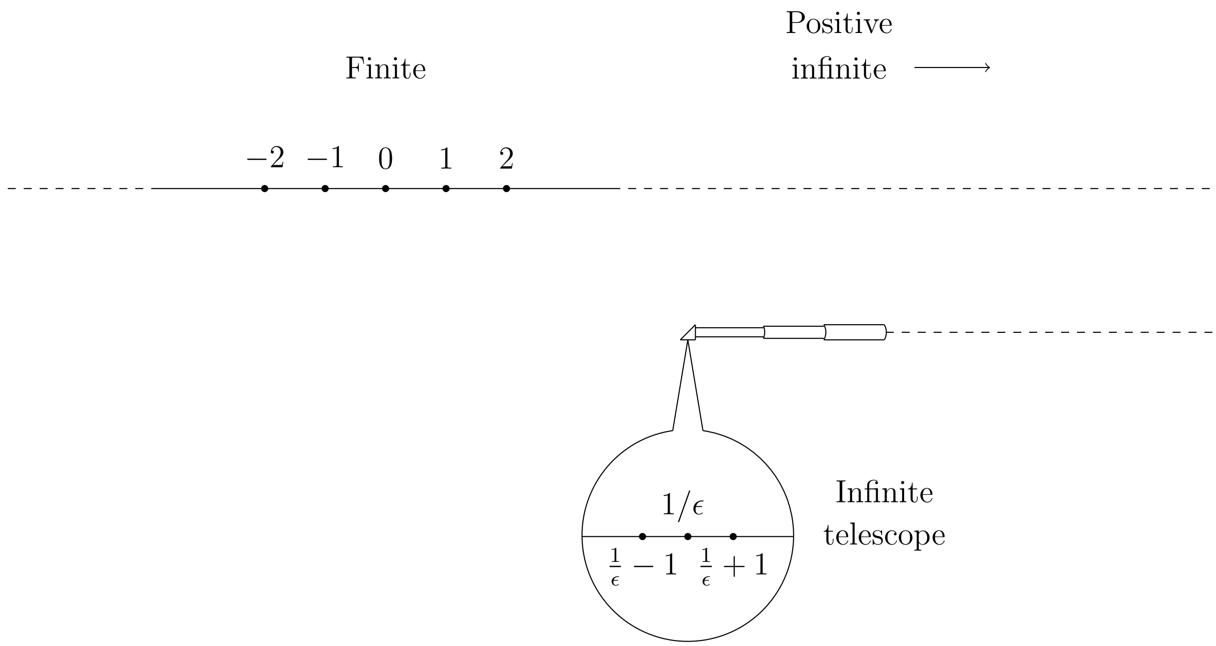

FIGURE 15.3.1. Keisler's telescope

which is the domain of the function called the *standard part function* (also known as the *shadow*) which rounds off each finite hyperreal to the nearest real number (for details see Section 15.6).

EXAMPLE 15.3.8. Slope calculation of $y = x^2$ at x_0. Here we use standard part function (also called the *shadow*)

$$\mathrm{st} : \{\text{finite hyperreals}\} \to \mathbb{R}$$

(for details concerning the *shadow* see Section 15.6). If a curve is defined by $y = x^2$ we wish to find the slope at the point x_0. To this end we use an infinitesimal x-increment Δx and compute the corresponding y-increment

$$\Delta y = (x_0 + \Delta x)^2 - x_0^2 = (x_0 + \Delta x + x_0)(x + \Delta x - x_0) = (2x_0 + \Delta x)\Delta x.$$

The corresponding "average" slope is therefore

$$\frac{\Delta y}{\Delta x} = 2x_0 + \Delta x$$

which is infinitely close to $2x$, and we are naturally led to the definition of the slope at x_0 as the *shadow* of $\frac{\Delta y}{\Delta x}$, namely $\mathrm{st}\left(\frac{\Delta y}{\Delta x}\right) = 2x_0$.

The extension principle expresses the idea that all real objects have natural hyperreal counterparts. We will be mainly interested in *sets*, *functions* and *relations*. We then have the following *extension principle*.

Extension principle. The order relation on the hyperreals contains the order relation on the reals. There exists a hyperreal number greater than zero but smaller than every positive real. Every set $D \subseteq \mathbb{R}$ has a natural extension ${}^*D \subseteq {}^*\mathbb{R}$. Every real function f with domain D has a natural hyperreal extension ${}^*f$ with domain ${}^*D$.[4]

Here the *naturality* of the extension alludes to the fact that such an extension is unique, and the *coherence* refers to the fact that the domain of the natural extension of a function is the natural extension of its domain.

A positive infinitesimal is a positive hyperreal smaller than every positive real. A negative infinitesimal is a negative hyperreal greater than every negative real. An arbitrary infinitesimal is either a positive infinitesimal, a negative infinitesimal, or zero.

Ultimately it turns out counterproductive to employ asterisks for hyperreal functions (in fact we already dropped it in equation (15.4.1)).

15.4. Transfer principle

DEFINITION 15.4.1. The Transfer Principle asserts that every first-order statement true over $\mathbb{R}$ is similarly true over ${}^*\mathbb{R}$, and vice versa.

Here the adjective *first-order* alludes to the limitation on quantification to elements as opposed to sets.

Listed below are a few examples of first-order statements.

EXAMPLE 15.4.2. The commutativity rule for addition $x+y = y+x$ is valid for all hyperreal x, y by the transfer principle.

EXAMPLE 15.4.3. The formula

$$\sin^2 x + \cos^2 x = 1 \tag{15.4.1}$$

is valid for all hyperreal x by the transfer principle.

EXAMPLE 15.4.4. The statement

$$0 < x < y \implies 0 < \frac{1}{y} < \frac{1}{x} \tag{15.4.2}$$

holds for all hyperreal x, y.

[4] Here the noun *principle* (in *extension principle*) means that we are going to assume that there is a function ${}^*f : \mathbb{B} \to \mathbb{B}$ which satisfies certain properties. It is a separate problem to define $\mathbb{B}$ which admits a coherent definition of ${}^*f$ for all $f : \mathbb{A} \to \mathbb{A}$, to be solved below.

EXAMPLE 15.4.5. The characteristic function $\chi_{\mathbb{Q}}$ of the rational numbers equals 1 on rational inputs and 0 on irrational inputs. By the transfer principle, its natural extension ${}^*\chi_{\mathbb{Q}} = \chi_{{}^*\mathbb{Q}}$ will be 1 on hyperrational numbers ${}^*\mathbb{Q}$ and 0 on hyperirrational numbers (namely, numbers in the complement ${}^*\mathbb{R} \setminus {}^*\mathbb{Q}$).

To give additional examples of real statements to which transfer applies, note that all *ordered field*-statements are subject to Transfer. As we will see below, it is possible to extend Transfer to a much broader category of statements, such as those containing the function symbols exp or sin or those that involve infinite sequences of reals.

A hyperreal number x is finite if there exists a real number r such that $|x| < r$. A hyperreal number is called *positive infinite* if it is greater than every real number, and *negative infinite* if it is smaller than every real number.

15.5. Orders of magnitude

Hyperreal numbers come in three orders of magnitude: infinitesimal, appreciable, and infinite. A number is appreciable if it is finite but not infinitesimal. In this section we will outline the rules for manipulating hyperreal numbers.

To give a typical proof, consider the rule that if ϵ is positive infinitesimal then $\frac{1}{\epsilon}$ is positive infinite. Indeed, for every positive real r we have $0 < \epsilon < r$. It follows from (15.4.2) by transfer that that $\frac{1}{\epsilon}$ is greater than every positive real, i.e., that $\frac{1}{\epsilon}$ is infinite.

Let ϵ, δ denote arbitrary infinitesimals. Let b, c denote arbitrary appreciable numbers. Let H, K denote arbitrary infinite numbers. We have the following theorem.

THEOREM 15.5.1. *We have the following rules for addition:*

- $\epsilon + \delta$ *is infinitesimal;*
- $b + \epsilon$ *is appreciable;*
- $b + c$ *is finite (possibly infinitesimal);*
- $H + \epsilon$ *and* $H + b$ *are infinite.*

We have the following rules for products.

- $\epsilon\delta$ *and* $b\epsilon$ *are infinitesimal;*
- bc *is appreciable;*
- Hb *and* HK *are infinite.*

We have the following rules for quotients.

- $\frac{\epsilon}{b}$, $\frac{\epsilon}{H}$, $\frac{b}{H}$ *are infinitesimal;*
- $\frac{b}{c}$ *is appreciable;*

- $\frac{b}{\epsilon}$, $\frac{H}{\epsilon}$, $\frac{H}{b}$ *are infinite provided* $\epsilon \neq 0$.

We have the following rules for roots, where n *is a standard natural number.*

- *if* $\epsilon > 0$ *then* $\sqrt[n]{\epsilon}$ *is infinitesimal;*
- *if* $b > 0$ *then* $\sqrt[n]{b}$ *is appreciable;*
- *if* $H > 0$ *then* $\sqrt[n]{H}$ *is infinite.*

Note that the traditional topic of the so-called "indeterminate forms" can be treated without introducing any ad-hoc terminology by means of the following remark.

Remark 15.5.2. We have *no rules* in certain cases, such as $\frac{\epsilon}{\delta}$, $\frac{H}{K}$, $H\epsilon$, and $H+K$.

These cases correspond to what are known since Moigno[5] as *indeterminate forms.*

Theorem 15.5.3. *Arithmetic operations on the hyperreal numbers are governed by the following rules.*

(1) *every hyperreal number between two infinitesimals is infinitesimal.*
(2) *Every hyperreal number which is between two finite hyperreal numbers, is finite.*
(3) *Every hyperreal number which is greater than some positive infinite number, is positive infinite.*
(4) *Every hyperreal number which is less than some negative infinite number, is negative infinite.*

Example 15.5.4. The difference $\sqrt{H+1} - \sqrt{H-1}$ (where H is infinite) is infinitesimal. Namely,

$$\begin{aligned}\sqrt{H+1} - \sqrt{H-1} &= \frac{\left(\sqrt{H+1} - \sqrt{H-1}\right)\left(\sqrt{H+1} + \sqrt{H-1}\right)}{\left(\sqrt{H+1} + \sqrt{H-1}\right)} \\ &= \frac{H+1-(H-1)}{\left(\sqrt{H+1} + \sqrt{H-1}\right)} \\ &= \frac{2}{\sqrt{H+1} + \sqrt{H-1}}\end{aligned}$$

is infinitesimal. Once we introduce limits, this example can be reformulated as follows: $\lim_{n\to\infty}\left(\sqrt{n+1} - \sqrt{n-1}\right) = 0$.

[5]Elaborate on this historical note.

Definition 15.5.5. Two hyperreal numbers a, b are said to be infinitely close, written

$$a \approx b,$$

if their difference $a - b$ is infinitesimal.

It is convenient also to introduce the following terminology and notation.

Definition 15.5.6. Two nonzero hyperreal numbers a, b are said to be *adequal*, written

$$a \sqcap b,$$

if either $\frac{a}{b} \approx 1$ or $a = b = 0$.

Note that the relation $\sin x \approx x$ for infinitesimal x is immediate from the continuity of sine at the origin (in fact both sides are infinitely close to 0), whereas the relation $\sin x \sqcap x$ is a subtler relation equivalent to the computation of the first order Taylor approximation of sine.

15.6. Standard part principle

Theorem 15.6.1 (Standard Part Principle). *Every finite hyperreal number x is infinitely close to an appropriate real number.*

Proof. The result is generally true for an arbitrary proper ordered field extension E of $\mathbb{R}$. Indeed, if $x \in E$ is finite, then x induces a Dedekind cut on the subfield $\mathbb{Q} \subseteq \mathbb{R} \subseteq E$ via the total order of E. The real number corresponding to the Dedekind cut is then infinitely close to x. □

The real number infinitely close to x is called the standard part, or *shadow*, denoted $\mathrm{st}(x)$, of x.

We will use the notation Δx, Δy for infinitesimals.

Remark 15.6.2. There are three consecutive stages in a typical calculation: (1) calculations with hyperreal numbers, (2) calculation with standard part, (3) calculation with real numbers.

15.7. Differentiation

An infinitesimal increment Δx can be visualized graphically by means of a microscope as in the Figure 15.7.1.

The slope s of a function f at a real point a is defined by setting

$$s = \mathrm{st}\left(\frac{f(a+\Delta x) - f(a)}{\Delta x}\right)$$

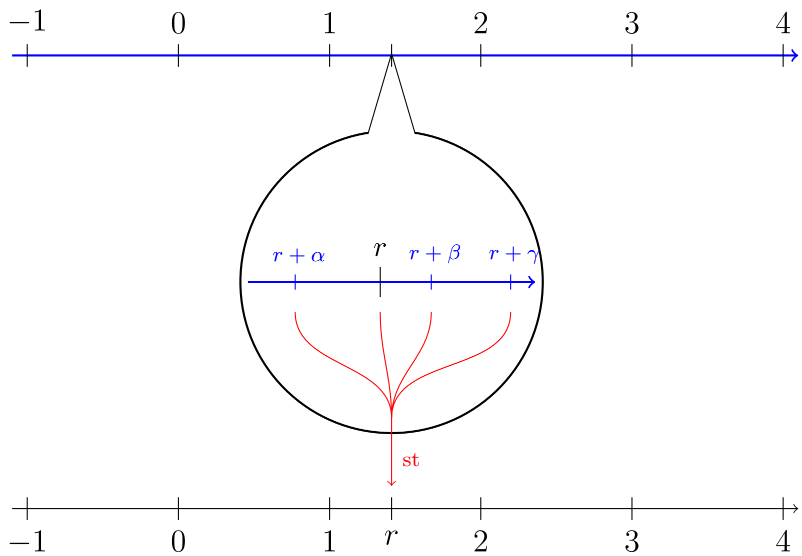

FIGURE 15.6.1. The standard part function, st, "rounds off" a finite hyperreal to the nearest real number. The function st is here represented by a vertical projection. Keisler's "infinitesimal microscope" is used to view an infinitesimal neighborhood of a standard real number r, where α, β, and γ represent typical infinitesimals. Courtesy of Wikipedia.

whenever the shadow exists (i.e., the ratio is finite) and is the same for each nonzero infinitesimal Δx. The construction is illustrated in Figure 15.7.2.

DEFINITION 15.7.1. Let f be a real function of one real variable. The *derivative* of f is the new function f' whose value at a real x is the slope of f at x. In symbols,

$$f'(x) = \operatorname{st}\left(\frac{f(x+\Delta x) - f(x)}{\Delta x}\right)$$

whenever the slope exists.

Equivalently, we can write $f'(x) \approx \frac{f(x+\Delta x)-f(x)}{\Delta x}$. When $y = f(x)$ we define a new dependent variable Δy by settting

$$\Delta y = f(x + \Delta x) - f(x)$$

called the y-increment, so we can write the derivative as $\operatorname{st}\left(\frac{\Delta y}{\Delta x}\right)$.

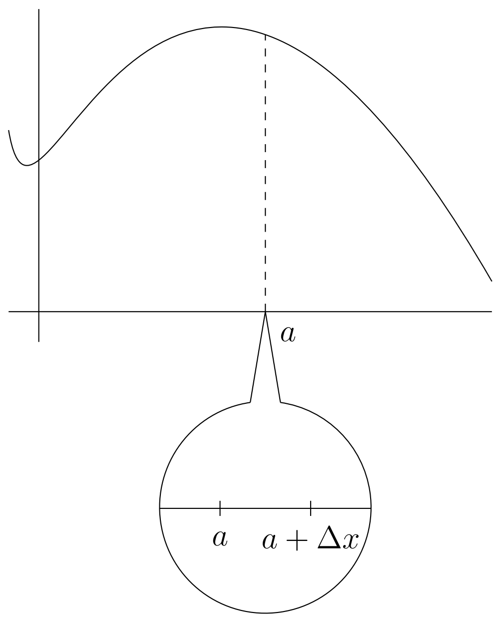

FIGURE 15.7.1. Infinitesimal increment Δx under the microscope

EXAMPLE 15.7.2. If $f(x) = x^2$ we obtain the derivative of $y = f(x)$ by the following direct calculation:

$$
\begin{aligned}
f'(x) &\approx \frac{\Delta y}{\Delta x} \\
&= \frac{(x + \Delta x)^2 - x^2}{\Delta x} \\
&= \frac{(x + \Delta x - x)(x + \Delta x + x)}{\Delta x} \\
&= \frac{\Delta x(2x + \Delta x)}{\Delta x} \\
&= 2x + \Delta x \\
&\approx 2x.
\end{aligned}
$$

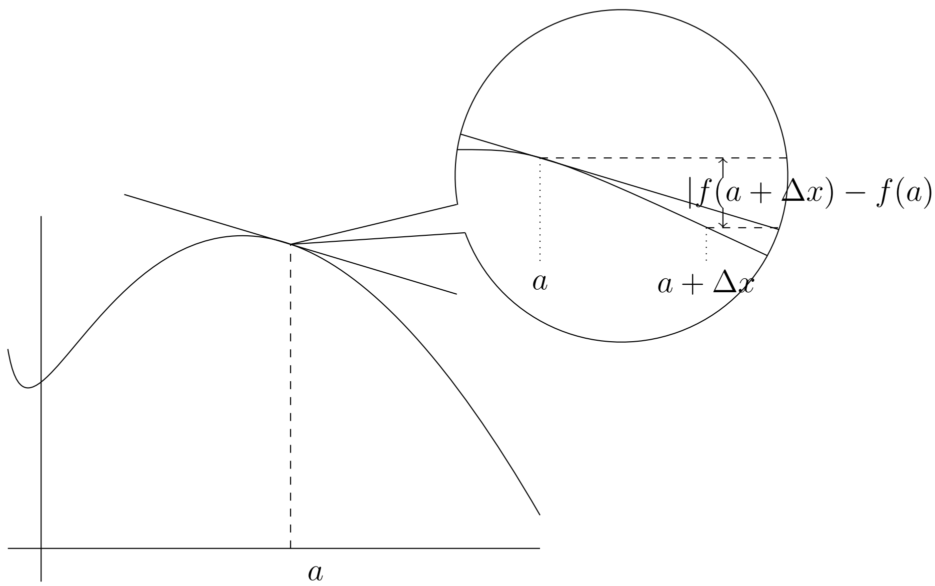

FIGURE 15.7.2. Defining slope of f at a

Given a function $y = f(x)$ one defines the dependent variable $\Delta y = f(x+\Delta x) - f(x)$ as above. One also defines a new dependent variable dy by setting $dy = f'(x)\Delta x$ at a point where f is differentiable, and sets for symmetry $dx = \Delta x$. Note that we have

$$dy \sqcap \Delta y$$

as in Definition 15.5.6 whenever $dy \neq 0$. We then have Leibniz's notation

$$\frac{dy}{dx}$$

for the derivative $f'(x)$, and rules like the chain rule acquire an appealing form.

CHAPTER 16

Global and systolic geometry

16.1. Definition of systole

The unit circle $S^1 \subset \mathbb{C}$ bounds the unit disk D. A loop on a surface M is a continuous map $S^1 \to M$.

Definition 16.1.1. A loop $S^1 \to M$ is called *contractible* if the map f extends from S^1 to the disk D by means of a continuous map $F : D \to M$.

Thus the restriction of F to S^1 is f.

The notions of contractible loops and simply connected spaces were reviewed in more detail in Section 17.1.

Definition 16.1.2. A loop is called noncontractible if it is not contractible.

Definition 16.1.3. Given a metric g on M, we will denote by $\mathrm{sys}_1(g)$, the infimum of lengths, referred to as the "systole" of g, of a noncontractible loop β in a compact, non-simply-connected Riemannian manifold (M, g):

$$\mathrm{sys}_1(g) = \inf_{\beta} \mathrm{length}(\beta), \tag{16.1.1}$$

where the infimum is over all noncontractible loops β in M. In graph theory, a similar invariant is known as the *girth* [**Tu47**].[1]

It can be shown that for a compact Riemannian manifold, the infimum is always attained, *cf.* Theorem 16.13.1. A loop realizing the minimum is necessarily a simple closed geodesic.

In systolic questions about surfaces, integral-geometric identities play a particularly important role. Roughly speaking, there is an integral identity relating area on the one hand, and an average of energies of a suitable family of loops, on the other. By the Cauchy-Schwarz inequality, there is an inequality relating energy and length squared, hence one obtains an inequality between area and the square of the systole.

[1]The notion of systole expressed by (16.1.1) is unrelated to the systolic arrays of [**Ku78**].

Such an approach works both for the Loewner inequality (16.2.1) and Pu's inequality (16.1.6) (biographical notes on C. Loewner and P. Pu appear in respectively). One can prove an inequality for the Möbius band this way, as well [**Bl61b**].

Here we prove the two classical results of systolic geometry, namely Loewner's torus inequality as well as Pu's inequality for the real projective plane.

16.1.1. Three systolic invariants. The material in this subsection is optional.

Let M be a Riemannian manifold. We define the homology 1-systole

$$\mathrm{sys}_1(M) \tag{16.1.2}$$

by minimizing $\mathrm{vol}(\alpha)$ over all nonzero homology classes. Namely, $\mathrm{sys}_1(M)$ is the least length of a loop C representing a nontrivial homology class $[C]$ in $H_1(M;\mathbb{Z})$.

We also define the stable homology systole

$$\mathrm{stsys}_1(M) = \lambda_1\left(H_1(M)/T_1,\ \|\ \|\right), \tag{16.1.3}$$

namely by minimizing the stable norm $\|\ \|$ of a class of infinite order (see Definition 17.13.1 for details).

Remark 16.1.4. For the real projective plane, these two systolic invariants are not the same. Namely, the homology systole sys_1 equals the least length of a noncontractible loop (which is also nontrivial homologically), while the stable systole is infinite being defined by a minimum over an empty set.

Recall the following example from the previous section:

Example 16.1.5. For an arbitrary metric on the 2-torus $\mathbb{T}^2$, the 1-systole and the stable 1-systole coincide by Theorem 17.5.3:

$$\mathrm{sys}_1(\mathbb{T}^2) = \mathrm{stsys}_1(\mathbb{T}^2),$$

for every metric on $\mathbb{T}^2$.

Using the notion of a noncontractible loop, we can define the homotopy 1-systole

$$\mathrm{sys}_1(M) \tag{16.1.4}$$

as the least length of a non-contractible loop in M.

In the case of the torus, the fundamental group $\mathbb{Z}^2$ is abelian and torsionfree, and therefore $\mathrm{sys}_1(\mathbb{T}^2) = \mathrm{sys}_1(\mathbb{T}^2)$, so that all three invariants coincide in this case.

16.1.2. Isoperimetric inequality and Pu's inequality. The material in this section is optional.

Pu's inequality can be thought of as an "opposite" isoperimetric inequality, in the following precise sense.

The classical isoperimetric inequality in the plane is a relation between two metric invariants: length L of a simple closed curve in the plane, and area A of the region bounded by the curve. Namely, every simple closed curve in the plane satisfies the inequality

$$\frac{A}{\pi} \leq \left(\frac{L}{2\pi}\right)^2.$$

This classical *isoperimetric inequality* is sharp, insofar as equality is attained only by a round circle.

In the 1950's, Charles Loewner's student P. M. Pu [**Pu52**] proved the following theorem. Let $\mathbb{RP}^2$ be the real projective plane endowed with an arbitrary metric, *i.e.* an embedding in some $\mathbb{R}^n$. Then

$$\left(\frac{L}{\pi}\right)^2 \leq \frac{A}{2\pi}, \tag{16.1.5}$$

where A is its total area and L is the length of its shortest non-contractible loop. This *isosystolic inequality*, or simply *systolic inequality* for short, is also sharp, to the extent that equality is attained only for a metric of constant Gaussian curvature, namely antipodal quotient of a round sphere, *cf.* Section 16.14. In our systolic notation (16.1.1), Pu's inequality takes the following form:

$$\operatorname{sys}_1(g)^2 \leq \tfrac{\pi}{2}\operatorname{area}(g), \tag{16.1.6}$$

for every metric g on $\mathbb{RP}^2$. See Theorem 16.16.2 for a discussion of the constant. The inequality is proved in `http://u.math.biu.ac.il/~katzmik/egreg826.pdf`

Pu's inequality can be generalized as follows. We will say that a surface is *aspherical* if it is not a 2-sphere.

THEOREM 16.1.6. *Every aspherical surface* (M, g) *satisfies the optimal bound* (16.1.6)*, attained precisely when, on the one hand, the surface* M *is a real projective plane, and on the other, the metric* g *is of constant Gaussian curvature.*

The extension to aspherical surfaces follows from Gromov's inequality (16.1.7) below (by comparing the numerical values of the two constants). Namely, every aspherical compact surface (M, g) admits a metric ball

$$B = B_p\left(\tfrac{1}{2}\operatorname{sys}_1(g)\right) \subseteq M$$

of radius $\frac{1}{2}\operatorname{sys}_1(g)$ which satisfies [**Gro83**, Corollary 5.2.B]

$$\operatorname{sys}_1(g)^2 \leq \frac{4}{3}\operatorname{area}(B). \tag{16.1.7}$$

16.1.3. Hermite and Bergé-Martinet constants. The material in this subsection is optional.

Most of the material in this section has already appeared in earlier chapters.

Let $b \in \mathbb{N}$. The Hermite constant γ_b is defined in one of the following two equivalent ways:

(1) γ_b is the *square* of the biggest first successive minimum, *cf.* Definition 16.17.1, among all lattices of unit covolume;
(2) γ_b is defined by the formula

$$\sqrt{\gamma_b} = \sup\left\{ \frac{\lambda_1(L)}{\operatorname{vol}(\mathbb{R}^b/L)^{1/b}} \,\middle|\, L \subseteq (\mathbb{R}^b, \|\ \|) \right\}, \tag{16.1.8}$$

where the supremum is extended over all lattices L in $\mathbb{R}^b$ with a Euclidean norm $\|\ \|$.

A lattice realizing the supremum is called a *critical* lattice. A critical lattice may be thought of as the one realizing the densest packing in $\mathbb{R}^b$ when we place balls of radius $\frac{1}{2}\lambda_1(L)$ at the points of L.

The existence of the Hermite constant, as well as the existence of critical lattices, are both nontrivial results [**Ca71**].

Theorem 14.9.1 provides the value for γ_2.

EXAMPLE 16.1.7. In dimensions $b \geq 3$, the Hermite constants are harder to compute, but explicit values (as well as the associated critical lattices) are known for small dimensions (≤ 8), *e.g.* $\gamma_3 = 2^{\frac{1}{3}} = 1.2599\ldots$, while $\gamma_4 = \sqrt{2} = 1.4142\ldots$. Note that γ_n is asymptotically linear in n, *cf.* (16.1.11).

A related constant γ'_b is defined as follows, *cf.* [**BeM**].

DEFINITION 16.1.8. The Bergé-Martinet constant γ'_b is defined by setting

$$\gamma'_b = \sup\left\{ \lambda_1(L)\lambda_1(L^*) \,\middle|\, L \subseteq (\mathbb{R}^b, \|\ \|) \right\}, \tag{16.1.9}$$

where the supremum is extended over all lattices L in $\mathbb{R}^b$.

Here L^* is the lattice dual to L. If L is the $\mathbb{Z}$-span of vectors (x_i), then L^* is the $\mathbb{Z}$-span of a dual basis (y_j) satisfying $\langle x_i, y_j\rangle = \delta_{ij}$, *cf.* relation (16.7.1).

Thus, the constant γ'_b is bounded above by the Hermite constant γ_b of (16.1.8). We have $\gamma'_1 = 1$, while for $b \geq 2$ we have the following inequality:

$$\gamma'_b \leq \gamma_b \leq \frac{2}{3}\, b \quad \text{for all } b \geq 2. \tag{16.1.10}$$

Moreover, one has the following asymptotic estimates:

$$\frac{b}{2\pi e}(1+o(1)) \leq \gamma'_b \leq \frac{b}{\pi e}\,(1+o(1)) \quad \text{for } b \to \infty, \tag{16.1.11}$$

cf. [**LaLS90**, pp. 334, 337]. Note that the lower bound of (16.1.11) for the Hermite constant and the Bergé-Martinet constant is nonconstructive, but see [**RT90**] and [**ConS99**].

Definition 16.1.9. A lattice L realizing the supremum in (16.1.9) or (16.1.9) is called *dual-critical*.

Remark 16.1.10. The constants γ'_b and the dual-critical lattices in $\mathbb{R}^b$ are explicitly known for $b \leq 4$, *cf.* [**BeM**, Proposition 2.13]. In particular, we have $\gamma'_1 = 1$, $\gamma'_2 = \frac{2}{\sqrt{3}}$.

Example 16.1.11. In dimension 3, the value of the Bergé-Martinet constant, $\gamma'_3 = \sqrt{\frac{3}{2}} = 1.2247\ldots$, is slightly below the Hermite constant $\gamma_3 = 2^{\frac{1}{3}} = 1.2599\ldots$. It is attained by the face-centered cubic lattice, which is not isodual [**MilH73**, p. 31], [**BeM**, Proposition 2.13(iii)], [**CoS94**].

This is the end of the three subsections containing optional material.

16.2. Loewner's torus inequality

Historically, the first lower bound for the volume of a Riemannian manifold in terms of a systole is due to Charles Loewner. In 1949, Loewner proved the first systolic inequality, in a course on Riemannian geometry at Syracuse University, *cf.* [**Pu52**]. Namely, he showed the following result, whose proof appears in Section 19.2.

Theorem 16.2.1 (C. Loewner). *Every Riemannian metric g on the torus $\mathbb{T}^2$ satisfies the inequality*

$$\operatorname{sys}_1(g)^2 \leq \gamma_2 \ \operatorname{area}(g), \tag{16.2.1}$$

where $\gamma_2 = \frac{2}{\sqrt{3}}$ is the Hermite constant (16.1.8). *A metric attaining the optimal bound* (16.2.1) *is necessarily flat, and is homothetic to the quotient of $\mathbb{C}$ by the Eisenstein integers, i.e. lattice spanned by the cube roots of unity, cf. Lemma 14.9.1.*

The result can be reformulated in a number of ways.[2] Loewner's torus inequality relates the total area, to the systole, *i.e.* least length of a noncontractible loop on the torus $(\mathbb{T}^2, g)$:

$$\operatorname{area}(g) - \tfrac{\sqrt{3}}{2}\operatorname{sys}_1(g)^2 \geq 0. \tag{16.2.2}$$

The boundary case of equality is attained if and only if the metric is homothetic to the flat metric obtained as the quotient of $\mathbb{R}^2$ by the lattice formed by the Eisenstein integers.

16.3. Loewner's inequality with remainder term

See `http://u.math.biu.ac.il/~katzmik/egreg826.pdf`

16.4. Global geometry of surfaces

Discussion of Local versus Global: The local behavior is by definition the behavior in an open neighborhood of a point. The local behavior of a smooth curve is well understood by the implicit function theorem. Namely, a smooth curve in the plane or in 3-space can be thought of as the graph of a smooth function. A curve in the plane is locally the graph of a scalar function. A curve in 3-space is locally the graph of a vector-valued function.

Example 16.4.1. The unit circle in the plane can be defined implicitly by

$$x^2 + y^2 = 1,$$

or parametrically by

$$t \mapsto (\cos t, \sin t).$$

Alternatively, it can be given locally as the graph of the function

$$f(x) = \sqrt{1 - x^2}.$$

Note that this presentation works only for points on the upper halfcircle. For points on the lower halfcircle we use the function

$$-\sqrt{1 - x^2}.$$

Both of these representations fail at the points $(1, 0)$ and $(-1, 0)$.

To overcome this difficulty, we must work with y as the independent variable, instead of x. Thus, we can parametrize a neighborhood of (1,0) by using the function

$$x = g(y) = \sqrt{1 - y^2}.$$

[2] Thus, in the case of the torus $\mathbb{T}^2$, the fundamental group is abelian. Hence the systole can be expressed in this case as follows: $\operatorname{sys}_1(\mathbb{T}^2) = \lambda_1\left(H_1(\mathbb{T}^2; \mathbb{Z}), \|\ \|\right)$, where $\|\ \|$ is the stable norm.

Example 16.4.2. The helix given in parametric form by

$$(x, y, z) = (\cos t, \sin t, t).$$

It can also be defined as the graph of the vector-valued function $f(z)$, with values in the (x, y)-plane, where

$$(x(z), y(z)) = f(z) = (\cos z, \sin z).$$

In this case the graph representation in fact works even globally.

Example 16.4.3. The unit sphere in 3-space can be represented locally as the graph of the function of two variables

$$f(x, y) = \sqrt{1 - x^2 - y^2}.$$

As above, concerning the local nature of the presentation necessitates additional functions to represent neighborhoods of points not in the open northern hemisphere (this example is discussed in more detail in Section 16.6).

16.5. Definition of manifold

Motivated by the examples given in the previous section, we give a general definition as follows.

A manifold is defined as a subset of Euclidean space which is locally a graph of a function, possibly vector-valued. This is the original definition of Poincaré who invented the notion (see Arnold [**1**, p. 234]).

Definition 16.5.1. By a 2-dimensional closed Riemannian manifold we mean a compact subset

$$M \subseteq \mathbb{R}^n$$

such that in an open neighborhood of every point $p \in M$ in $\mathbb{R}^n$, the compact subset M can be represented as the graph of a suitable smooth vector-valued function of two variables.

Here the function has values in $(n-2)$-dimensional vectors.

The usual parametrisation of the graph can then be used to calculate the coefficients g_{ij} of the first fundamental form, as, for example, in Theorem 16.16.2 and Definition 5.9.2. The collection of all such data is then denoted by the pair (M, g), where

$$g = (g_{ij})$$

is referred to as *the metric*.

Remark 16.5.2. Differential geometers like the (M, g) notation, because it helps separate the topology M from the geometry g. Strictly speaking, the notation is redundant, since the object g already incorporates all the information, including the topology. However, geometers have found it useful to use g when one wants to emphasize the geometry, and M when one wants to emphasize the topology.

Note that, as far as the intrinsic geometry of a Riemannian manifold is concerned, the embedding in $\mathbb{R}^n$ referred to in Definition 16.5.1 is irrelevant to a certain extent, all the more so since certain basic examples, such as flat tori, are difficult to imbed in a transparent way.

16.6. Sphere as a manifold

The round 2-sphere $S^2 \subseteq \mathbb{R}^3$ defined by the equation

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

is a closed Riemannian manifold. Indeed, consider the function $f(x, y)$ defined in the unit disk $x^2 + y^2 < 1$ by setting $f(x, y) = \sqrt{1 - x^2 - y^2}$. Define a coordinate chart

$$\underline{x}_1(u^1, u^2) = (u^1, u^2, f(u^1, u^2)).$$

Thus, each point of the open northern hemisphere admits a neighborhood diffeomorphic to a ball (and hence to $\mathbb{R}^2$). To cover the southern hemisphere, use the chart

$$x_2(u^1, u^2) = (u^1, u^2, -f(u^1, u^2)).$$

To cover the points on the equator, use in addition charts $\underline{x}_3(u^1, u^2) = (u^1, f(u^1, u^2), u^2)$, $\underline{x}_4(u^1, u^2) = (u^1, -f(u^1, u^2), u^2)$, as well as the pair of charts $\underline{x}_5(u^1, u^2) = (f(u^1, u^2), u^1, u^2)$, $\underline{x}_6(u^1, u^2) = (-f(u^1, u^2), u^1, u^2)$.

16.7. Dual bases

Tangent space, cotangent space, and the notation for bases in these spaces were discussed in Section 13.7.

We will work with dual bases $\left(\frac{\partial}{\partial u^i}\right)$ for vectors, and (du^i) for covectors (*i.e.* elements of the dual space), such that

$$du^i \left(\frac{\partial}{\partial u^j} \right) = \delta^i_j, \tag{16.7.1}$$

where δ^i_j is the Kronecker delta.

Recall that the metric coefficients are defined by setting

$$g_{ij} = \langle \underline{x}_i, \underline{x}_j \rangle,$$

where $\underline{x}$ is the parametrisation of the surface. We will only work with metrics whose first fundamental form is diagonal. We can thus write the first fundamental form as follows:

$$g = g_{11}(u^1, u^2)(du^1)^2 + g_{22}(u^1, u^2)(du^2)^2. \tag{16.7.2}$$

REMARK 16.7.1. Such data can be computed from a Euclidean embedding as usual, or it can be given apriori without an embedding, as we did in the case of the hyperlobic metric.

We will work with such data independently of any Euclidean embedding, as discussed in Section 10.7. For example, if the metric coefficients form an identity matrix, we obtain

$$g = \left(du^1\right)^2 + \left(du^2\right)^2, \tag{16.7.3}$$

where the interior superscript denotes an index, while exterior superscript denotes the squaring operation.

16.8. Jacobian matrix

The Jacobian matrix of $v = v(u)$ is the matrix

$$\frac{\partial(v^1, v^2)}{\partial(u^1, u^2)},$$

which is the matrix of partial derivatives. Denote by

$$\mathrm{Jac}_v(u) = \det\left(\frac{\partial(v^1, v^2)}{\partial(u^1, u^2)}\right).$$

It is shown in advanced calculus that for any function $f(v)$ in a domain D, one has

$$\int_D f(v)dv^1dv^2 = \int_D g(u)\mathrm{Jac}_v(u)du^1du^2, \tag{16.8.1}$$

where $g(u) = f(v(u))$.

EXAMPLE 16.8.1. Let $u^1 = r$, $u^2 = \theta$. Let $v^1 = x$, $v^2 = y$. We have $x = r\cos\theta$ and $y = r\sin\theta$. One easily shows that the Jacobian is $\mathrm{Jac}_v(u) = r$. The area elements are related by

$$dv^1dv^2 = \mathrm{Jac}_v(u)du^1du^2,$$

or

$$dxdy = rdrd\theta.$$

A similar relation holds for integrals.

16.9. Area of a surface, independence of partition

Partition[3] is what allows us to perform actual calculations with area, but the result is independent of partition (see below).

Based on the local definition of area discussed in an earlier chapter, we will now deal with the corresponding global invariant.

DEFINITION 16.9.1. The *area element* dA of the surface is the element

$$dA := \sqrt{\det(g_{ij})}du^1du^2,$$

where $\det(g_{ij}) = g_{11}g_{22} - g_{12}^2$ as usual.

THEOREM 16.9.2. *Define the* area *of* (M, g) *by means of the formula*

$$\text{area} = \int_M dA = \sum_{\{U\}} \int_U \sqrt{\det(g_{ij})}du^1du^2, \tag{16.9.1}$$

namely by choosing a partition $\{U\}$ *of* M *subordinate to a finite open cover as in Definition 16.5.1, performing a separate integration in each open set, and summing the resulting areas. Then the total area is independent of the partition and choice of coordinates.*

PROOF. Consider a change from a coordinate chart (u^i) to another coordinate chart, denoted (v^α). In the overlap of the two domains, the coordinates can be expressed in terms of each other, *e.g.* $v = v(u)$, and we have the 2 by 2 Jacobian matrix $\mathrm{Jac}_v(u)$.

Denote by $\tilde{g}_{\alpha\beta}$ the metric coefficients with respect to the chart (v^α). Thus, in the case of a metric induced by a Euclidean embedding defined by $x = x(u) = x(u^1, u^2)$, we obtain a new parametrisation

$$y(v) = x(u(v)).$$

Then we have

$$\begin{aligned}
\tilde{g}_{\alpha\beta} &= \left\langle \frac{\partial y}{\partial v^\alpha}, \frac{\partial y}{\partial v^\beta} \right\rangle \\
&= \left\langle \frac{\partial x}{\partial u^i}\frac{\partial u^i}{\partial v^\alpha}, \frac{\partial x}{\partial u^j}\frac{\partial u^j}{\partial v^\beta} \right\rangle \\
&= \frac{\partial u^i}{\partial v^\alpha}\frac{\partial u^j}{\partial v^\beta} \left\langle \frac{\partial x}{\partial u^i}, \frac{\partial x}{\partial u^j} \right\rangle \\
&= g_{ij}\frac{\partial u^i}{\partial v^\alpha}\frac{\partial u^j}{\partial v^\beta}.
\end{aligned}$$

[3]ritzuf or chaluka?

The right hand side is a product of *three* square matrices:

$$\frac{\partial u^i}{\partial v^\alpha}\, g_{ij}\, \frac{\partial u^j}{\partial v^\beta}.$$

The matrices on the left and on the right are both Jacobian matrices. Since determinant is multiplicative, we obtain

$$\det(\tilde{g}_{\alpha\beta}) = \det(g_{ij})\det\left(\frac{\partial(u^1,u^2)}{\partial(v^1,v^2)}\right)^2.$$

Hence using equation (16.8.1), we can write the area element as

$$\begin{aligned} dA &= \det^{\frac{1}{2}}(\tilde{g}_{\alpha\beta}) dv^1 dv^2 \\ &= \det^{\frac{1}{2}}(\tilde{g}_{\alpha\beta})\ \mathrm{Jac}_v(u) du^1 du^2 \\ &= \det^{\frac{1}{2}}(g_{ij})\ \det\left(\frac{\partial(u^1,u^2)}{\partial(v^1,v^2)}\right) \mathrm{Jac}_v(u) du^1 du^2 \\ &= \det^{\frac{1}{2}}(g_{ij})\ du^1 du^2 \end{aligned}$$

since inverse maps have reciprocal Jacobians by chain rule. Thus the integrand is unchanged and the area element is well defined. □

16.10. Conformal equivalence

Definition 16.10.1. Two metrics, $g = g_{ij} du^i du^j$ and $h = h_{ij} du^i du^j$, on M are called *conformally equivalent*, or *conformal* for short, if there exists a function $f = f(u^1,u^2) > 0$ such that

$$g = f^2 h,$$

in other words,

$$g_{ij} = f^2 h_{ij} \quad \forall i,j. \tag{16.10.1}$$

Definition 16.10.2. The function f above is called the *conformal factor* (note that sometimes it is more convenient to refer, instead, to the function $\lambda = f^2$ as the conformal factor).

Theorem 16.10.3. *Note that the length of every vector at a given point* (u^1,u^2) *is multiplied precisely by* $f(u^1,u^2)$.

Proof. More speficially, a vector $v = v^i \frac{\partial}{\partial u^i}$ which is a unit vector for the metric h, is "stretched" by a factor of f, *i.e.* its length with respect to g equals f. Indeed, the new length of v is

$$\begin{aligned}\sqrt{g(v,v)} &= g\left(v^i\frac{\partial}{\partial u^i}, v^j\frac{\partial}{\partial u^j}\right)^{\frac{1}{2}} \\ &= \left(g\left(\frac{\partial}{\partial u^i}, \frac{\partial}{\partial u^j}\right)v^iv^j\right)^{\frac{1}{2}} \\ &= \sqrt{g_{ij}v^iv^j} \\ &= \sqrt{f^2h_{ij}v^iv^j} \\ &= f\sqrt{h_{ij}v^iv^j} \\ &= f,\end{aligned}$$

proving the theorem. □

Definition 16.10.4. An equivalence class of metrics on M conformal to each other is called a *conformal structure* on M (mivneh conformi).

16.11. Geodesic equation

The material in this section has already been dealt with in an earlier chapter.

Perhaps the simplest possible definition of a geodesic β on a surface in 3-space is in terms of the orthogonality of its second derivative β'' to the surface. The nonlinear second order ordinary differential equation defining a geodesic is, of course, the "true" if complicated definition. We will now prove the equivalence of the two definitions. Consider a plane curve

$$\underset{s}{\mathbb{R}} \underset{\alpha}{\longrightarrow} \underset{(u^1,u^2)}{\mathbb{R}^2}$$

where $\alpha = (\alpha^1(s), \alpha^2(s))$. Let $\underline{x}\colon \mathbb{R}^2 \to \mathbb{R}^3$ be a regular parametrisation of a surface in 3-space. Then the composition

$$\underset{s}{\mathbb{R}} \underset{\alpha}{\longrightarrow} \underset{(u^1,u^2)}{\mathbb{R}^2} \underset{x}{\longrightarrow} \mathbb{R}^3$$

yields a curve

$$\beta = x \circ \alpha.$$

Definition 16.11.1. A curve $\beta = x\circ\alpha$ is a geodesic on the surface x if one of the following two equivalent conditions is satisfied:

(a) we have for each $k = 1, 2$,

$$(\alpha^k)'' + \Gamma^k_{ij}(\alpha^i)'(\alpha^j)' = 0 \quad \text{where} \quad {}' = \frac{d}{ds}\ , \tag{16.11.1}$$

meaning that

$$(\forall k) \quad \frac{d^2\alpha^k}{ds^2} + \Gamma^k_{ij}\frac{d\alpha^i}{ds}\frac{d\alpha^j}{ds} = 0;$$

(b) the vector β'' is perpendicular to the surface and one has

$$\beta'' = L_{ij}\alpha^{i\prime}\alpha^{j\prime}n. \tag{16.11.2}$$

To prove the equivalence, we write $\beta = x \circ \alpha$, then $\beta' = x_i\alpha^{i\prime}$ by chain rule. Furthermore,

$$\beta'' = \frac{d}{ds}(x_i \circ \alpha)\alpha^{i\prime} + x_i\alpha^{i\prime\prime} = x_{ij}\alpha^{j\prime}\alpha^{i\prime} + x_k\alpha^{k\prime\prime}.$$

Since $x_{ij} = \Gamma^k_{ij}x_k + L_{ij}n$ holds, we have

$$\beta'' - L_{ij}\alpha^{i\prime}\alpha^{j\prime}n = x_k\left(\alpha^{k\prime\prime} + \Gamma^k_{ij}\alpha^{i\prime}\alpha^{j\prime}\right).$$

16.12. Closed geodesic

DEFINITION 16.12.1. A *closed geodesic* in a Riemannian 2-manifold M is defined equivalently as

(1) a periodic curve $\beta : \mathbb{R} \to M$ satisfying the geodesic equation $\alpha^{k\prime\prime} + \Gamma^k_{ij}\alpha^{i\prime}\alpha^{j\prime} = 0$ in every chart $\underline{x} : \mathbb{R}^2 \to M$, where, as usual, $\beta = \underline{x} \circ \alpha$ and $\alpha(s) = (\alpha^1(s), \alpha^2(s))$ where s is arclength. Namely, there exists a period $T > 0$ such that $\beta(s+T) = \beta(s)$ for all s.
(2) A unit speed map from a circle $\mathbb{R}/L_T \to M$ satisfying the geodesic equation at each point, where $L_T = T\mathbb{Z} \subseteq \mathbb{R}$ is the rank one lattice generated by $T > 0$.

DEFINITION 16.12.2. The *length* $L(\beta)$ of a path $\beta : [a, b] \to M$ is calculated using the formula

$$L(\beta) = \int_a^b \|\beta'(t)\| dt,$$

where $\|v\| = \sqrt{g_{ij}v^iv^j}$ whenever $v = v^i\underline{x}_i$. The energy is defined by $E(\beta) = \int_a^b \|\beta'(t)\|^2 dt$.

A closed geodesic as in Definition 16.12.1, item 2 has length T.

REMARK 16.12.3. The geodesic equation (16.11.1) is the Euler-Lagrange equation of the first variation of arc length. Therefore when a path minimizes arc length among all neighboring paths connecting two fixed points, it must be a geodesic. A corresponding statement is valid for closed loops, *cf.* proof of Theorem 16.13.1. See also Section 16.1.

In Sections 16.14 and 16.18 we will give a complete description of the geodesics for the constant curvature sphere, as well as for flat tori.

16.13. Existence of closed geodesic

THEOREM 16.13.1. *Every free homotopy class of loops in a closed manifold contains a closed geodesic.*

PROOF. We sketch a proof for the benefit of a curious reader, who can also check that the construction is independent of the choices involved. The relevant topological notions are defined in Section 17.1 and [**Hat02**]. A free homotopy class α of a manifold M corresponds to a conjugacy class $g_\alpha \subset \pi_1(M)$. Pick an element $g \in g_\alpha$. Thus g acts on the universal cover $\tilde{M}$ of M. Let $f_g : M \to \mathbb{R}$ be the displacement function of g, *i.e.*

$$f_g(x) = d(\tilde{x}, g.\tilde{x}).$$

Let $x_0 \in M$ be a minimum of f_g. A first variation argument shows that any length-minimizing path between $\tilde{x}_0$ and $g.\tilde{x}_0$ descends to a closed geodesic in M representing α, *cf.* [**Car92, Ch93, GaHL04**]. □

16.14. Surfaces of constant curvature

By the uniformisation theorem 13.12.2, all surfaces fall into three types, according to whether they are conformally equivalent to metrics that are:

(1) flat (*i.e.* have zero Gaussian curvature $K \equiv 0$);
(2) spherical ($K \equiv +1$);
(3) hyperbolic ($K \equiv -1$).

For closed surfaces, the sign of the Gaussian curvature K is that of its Euler characteristic, *cf.* formula (17.9.1).

THEOREM 16.14.1 (Constant positive curvature). *There are only two compact surfaces, up to isometry, of constant Gaussian curvature $K = +1$. They are the round sphere S^2 of Example 16.6; and the real projective plane, denoted $\mathbb{RP}^2$.*

16.15. Real projective plane

Intuitively, one thinks of the real projective plane as the quotient surface obtained if one starts with the northern hemisphere of the 2-sphere, and "glues" together pairs of opposite points of the equatorial circle (the boundary of the hemisphere).

More formally, the real projective plane can be defined as follows. Let

$$m : S^2 \to S^2$$

denote the antipodal map of the sphere, *i.e.* the restriction of the map

$$v \mapsto -v$$

in $\mathbb{R}^3$. Then m is an involution. In other words, if we consider the action of the group $\mathbb{Z}_2 = \{e, m\}$ on the sphere, each orbit of the $\mathbb{Z}_2$ action on S^2 consists of a pair of antipodal points

$$\{\pm p\} \subseteq S^2. \tag{16.15.1}$$

DEFINITION 16.15.1. On the set-theoretic level, the real projective plane $\mathbb{RP}^2$ is the set of orbits of type (16.15.1), *i.e.* the quotient of S^2 by the $\mathbb{Z}_2$ action.

Denote by $Q : S^2 \to \mathbb{RP}^2$ the quotient map. The smooth structure and metric on $\mathbb{RP}^2$ are induced from S^2 in the following sense. Let $\underline{x} : \mathbb{R}^2 \to S^2$ be a chart on S^2 not containing any pair of antipodal points. Let g_{ij} be the metric coefficients with respect to this chart.

Let $\underline{y} = m \circ \underline{x} = -\underline{x}$ denote the "opposite" chart, and denote by h_{ij} its metric coefficients. Then

$$h_{ij} = \langle \frac{\partial \underline{y}}{\partial u^i}, \frac{\partial \underline{y}}{\partial u^j} \rangle = \langle -\frac{\partial \underline{x}}{\partial u^i}, -\frac{\partial \underline{x}}{\partial u^j} \rangle = \langle \frac{\partial \underline{x}}{\partial u^i}, \frac{\partial \underline{x}}{\partial u^j} \rangle = g_{ij}. \tag{16.15.2}$$

Thus the opposite chart defines the identical metric coefficients. The composition $Q \circ \underline{x}$ is a chart on $\mathbb{RP}^2$, and the same functions g_{ij} form the metric coefficients for $\mathbb{RP}^2$ relative to this chart.

We can summarize the preceeding discussion by means of the following definition.

DEFINITION 16.15.2. The real projective plane $\mathbb{RP}^2$ is defined in the following two equivalent ways:

(1) the quotient of the round sphere S^2 by (the restriction to S^2 of) the antipodal map $v \mapsto -v$ in $\mathbb{R}^3$. In other words, a typical point of $\mathbb{RP}^2$ can be thought of as a pair of opposite points of the round sphere.
(2) the northern hemisphere of S^2, with opposite points of the equator identified.

The smooth structure of $\mathbb{RP}^2$ is induced from the round sphere. Since the antipodal map preserves the metric coefficients by the calculation (16.15.2), the metric structure of constant Gaussian curvature $K = +1$ descends to $\mathbb{RP}^2$, as well.

16.16. Simple loops for surfaces of positive curvature

DEFINITION 16.16.1. A loop $\alpha : S^1 \to X$ of a space X is called *simple* if the map α is one-to-one, *cf.* Definition 17.1.1.

THEOREM 16.16.2. *The basic properties of the geodesics on surfaces of constant positive curvature as as follows:*

(1) *all geodesics are closed;*
(2) *the simple closed geodesics on S^2 have length 2π and are defined by the great circles;*
(3) *the simple closed geodesics on $\mathbb{RP}^2$ have length π;*
(4) *the simple closed geodesics of $\mathbb{RP}^2$ are parametrized by half-great circles on the sphere.*

PROOF. We calculate the length of the equator of S^2. Here the sphere $\rho = 1$ is parametrized by

$$x(\theta, \varphi) = (\sin\varphi\cos\theta, \sin\varphi\sin\theta, \cos\varphi)$$

in spherical coordinates (θ, φ). The equator is the curve $x \circ \alpha$ where $\alpha(s) = (s, \frac{\pi}{2})$ with $s \in [0, 2\pi]$. Thus $\alpha^1(s) = \theta(s) = s$. Recall that the metric coefficients are given by $g_{11}(\theta, \varphi) = \sin^2\varphi$, while $g_{22} = 1$ and $g_{12} = 0$. Thus

$$\|\beta'(s)\| = \sqrt{g_{ij}\left(s, \tfrac{\pi}{2}\right)\alpha^{i\prime}\alpha^{j\prime}} = \sqrt{\left(\sin\tfrac{\pi}{2}\right)\left(\tfrac{d\theta}{ds}\right)^2} = 1.$$

Thus the length of β is

$$\int_0^{2\pi} \|\beta'(s)\| ds = \int_0^{2\pi} 1 ds = 2\pi.$$

A geodesic on $\mathbb{RP}^2$ is twice as short as on S^2, since the antipodal points are identified, and therefore the geodesic "closes up" sooner than (*i.e.* twice as fast as) on the sphere. For example, a longitude of S^2 is not a closed curve, but it descends to a closed curve on $\mathbb{RP}^2$, since its endpoints (north and south poles) are antipodal, and are therefore identified with each other. $\square$

16.17. Successive minima

The material in this section has already been dealt with in an earlier chapter.

Let B be a finite-dimensional Banach space, *i.e.* a vector space together with a norm $\| \, \|$. Let $L \subseteq (B, \| \, \|)$ be a lattice of maximal rank, *i.e.* satisfying $\operatorname{rank}(L) = \dim(B)$. We define the notion of successive minima of L as follows, *cf.* [**GruL87**, p. 58].

DEFINITION 16.17.1. For each $k = 1, 2, \ldots, \operatorname{rank}(L)$, define the *$k$-th successive minimum* λ_k of the lattice L by

$$\lambda_k(L, \| \, \|) = \inf\left\{\lambda \in \mathbb{R} \;\middle|\; \begin{array}{l} \exists \text{ lin. indep. } v_1, \ldots, v_k \in L \\ \text{with } \|v_i\| \le \lambda \text{ for all } i \end{array}\right\}. \quad (16.17.1)$$

Thus the first successive minimum, $\lambda_1(L, \| \|)$ is the least length of a nonzero vector in L.

16.18. Flat surfaces

A metric is called *flat* if its Gaussian curvature K vanishes at every point.

THEOREM 16.18.1. *A closed surface of constant Gaussian curvature $K = 0$ is topologically either a torus $\mathbb{T}^2$ or a Klein bottle.*

Let us give a precise description in the former case.

EXAMPLE 16.18.2 (Flat tori). Every flat torus is isometric to a quotient $\mathbb{T}^2 = \mathbb{R}^2/L$ where L is a lattice, *cf.* [**Lo71**, Theorem 38.2]. In other words, a point of the torus is a coset of the additive action of the lattice in $\mathbb{R}^2$. The smooth structure is inherited from $\mathbb{R}^2$. Meanwhile, the additive action of the lattice is isometric. Indeed, we have $\mathrm{dist}(p, q) = \|q - p\|$, while for any $\ell \in L$, we have

$$\mathrm{dist}(p + \ell, q + \ell) = \|q + \ell - (p + \ell)\| = \|q - p\| = \mathrm{dist}(p, q).$$

Therefore the flat metric on $\mathbb{R}^2$ descends to $\mathbb{T}^2$.

Note that locally, all flat tori are indistinguishable from the flat plane itself. However, their global geometry depends on the metric invariants of the lattice, *e.g.* its successive minima, *cf.* Definition 16.17.1. Thus, we have the following.

THEOREM 16.18.3. *The least length of a nontrivial closed geodesic on a flat torus $\mathbb{T}^2 = \mathbb{R}^2/L$ equals the first successive minimum $\lambda_1(L)$.*

PROOF. The geodesics on the torus are the projections of straight lines in $\mathbb{R}^2$. In order for a straight line to close up, it must pass through a pair of points x and $x+\ell$ where $\ell \in L$. The length of the corresponding closed geodesic on $\mathbb{T}^2$ is precisely $\|\ell\|$, where $\| \|$ is the Euclidean norm. By choosing a shortest element in the lattice, we obtain a shortest closed geodesic on the corresponding torus. □

16.19. Hyperbolic surfaces

Most closed surfaces admit neither flat metrics nor metrics of positive curvature, but rather hyperbolic metrics. A hyperbolic surface is a surface equipped with a metric of constant Gaussian curvature $K = -1$. This case is far richer than the other two.

EXAMPLE 16.19.1. The pseudosphere (so called because its Gaussian curvature is constant, and equals -1) is the surface of revolution

$$(f(\varphi)\cos\theta, f(\varphi)\sin\theta, g(\varphi))$$

in $\mathbb{R}^3$ defined by the functions $f(\phi) = e^\phi$ and

$$g(\phi) = \int_0^\phi \left(1 - e^{2\psi}\right)^{1/2} d\psi,$$

where ϕ ranges through the interval $-\infty < \phi \le 0$. The usual formulas $g_{11} = f^2$ as well as $g_{22} = \left(\frac{dr}{d\phi}\right)^2 + \left(\frac{dg}{d\phi}\right)^2$ yield in our case $g_{11} = e^{2\phi}$, while

$$\begin{aligned} g_{22} &= \left(e^\phi\right)^2 + \left(\sqrt{1 - e^{2\phi}}\right)^2 \\ &= e^{2\phi} + 1 - e^{2\phi} \\ &= 1. \end{aligned}$$

Thus $(g_{ij}) = \begin{pmatrix} e^{2\phi} & 0 \\ 0 & 1 \end{pmatrix}$. The pseudosphere has constant Gaussian curvature -1, but it is not a closed surface (as it is unbounded in $\mathbb{R}^3$).

16.20. Hyperbolic plane

This was already discussed in Section 10.2. The metric

$$g_{\mathcal{H}^2} = \frac{1}{y^2}(dx^2 + dy^2) \tag{16.20.1}$$

in the upper half-plane

$$\mathcal{H}^2 = \{(x, y) \mid y > 0\}$$

is called the hyperbolic metric of the upper half plane.

THEOREM 16.20.1. *The metric* (16.20.1) *has constant Gaussian curvature* $K = -1$.

PROOF. By Theorem 11.2.2, we have

$$K = -\Delta_{LB} \ln f = \Delta_{LB} \ln y = y^2 \left(-\frac{1}{y^2}\right) = -1,$$

as required. □

In coordinates (u^1, u^2), we can write it, a bit awkwardly, as

$$g_{\mathcal{H}^2} = \frac{1}{(u^2)^2}\left(\left(du^1\right)^2 + \left(du^2\right)^2\right).$$

The Riemannian manifold $(\mathcal{H}^2, g_{\mathcal{H}^2})$ is referred to as the *Poincaré upper half-plane*. Its significance resides in the following theorem.

THEOREM 16.20.2. *Every closed hyperbolic surface M is isometric to the quotient of the Poincaré upper half-plane by the action of a suitable group Γ:*

$$M = \mathcal{H}^2/\Gamma.$$

Here the nonabelian group Γ is a discrete subgroup $\Gamma \subseteq PSL(2,\mathbb{R})$, where a matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ acts on $\mathcal{H}^2 = \{z \in \mathbb{C} | \Im(z) > 0\}$ by

$$z \mapsto \frac{az+b}{cz+d},$$

called fractional linear transformations, of Mobius transformations. All such transformations are isometries of the hyperbolic metric. The following theorem is proved, for example, in [**Kato92**].

THEOREM 16.20.3. *Every geodesic in the Poincaré upper half-plane is either a vertical ray, or a semicircle perpendicular to the x-axis.*

The foundational significance of this model in the context of the parallel postulate of Euclid has been discussed by numerous authors.

EXAMPLE 16.20.4. The length of a vertical interval joining i to ci can be calculated as follows. Recall that the conformal factor is $f(x,y) = \frac{1}{y}$. The length is therefore given by

$$\left| \int_1^c \frac{1}{y} dy \right| = |\ln c|.$$

Here the substitution $y = e^s$ gives an arclength parametrisation.

CHAPTER 17

Elements of the topology of surfaces

17.1. Loops, simply connected spaces

We would like to provide a self-contained explanation of the topological ingredient which is necessary so as to understand Loewner's torus inequality, *i.e.* essentially the notion of a noncontractible loop and the fundamental group of a topological space X. See [**Hat02**, Chapter 1] for a more detailed account.

Definition 17.1.1. A *loop* in X can be defined in one of two equivalent ways:

(1) a continuous map $\beta : [a,b] \to X$ satisfying $\beta(a) = \beta(b)$;
(2) a continuous map $\lambda : S^1 \to X$ from the circle S^1 to X.

Lemma 17.1.2. *The two definitions of a loop are equivalent.*

Proof. Consider the unique increasing linear function

$$f : [a,b] \to [0, 2\pi]$$

which is one-to-one and onto. Thus, $f(t) = \frac{2\pi(t-a)}{b-a}$. Given a map

$$\lambda(e^{is}) : S^1 \to X,$$

we associate to it a map $\beta(t) = \lambda\left(e^{if(t)}\right)$, and vice versa. □

Definition 17.1.3. A loop $S^1 \to X$ is said to be *contractible* if the map of the circle can be extended to a continuous map of the unit disk $\mathbb{D} \to X$, where $S^1 = \partial\mathbb{D}$.

Definition 17.1.4. A space X is called *simply connected* if every loop in X is contractible.

Theorem 17.1.5. *The sphere $S^n \subseteq \mathbb{R}^{n+1}$ which is the solution set of $x_0^2 + \ldots + x_n^2 = 1$ is simply connected for $n \geq 2$. The circle S^1 is not simply connected.*

17.2. Orientation on loops and surfaces

Let $S^1 \subseteq \mathbb{C}$ be the unit circle. The choice of an orientation on the circle is an arrow pointing clockwise or counterclockwise. The standard choice is to consider S^1 as an oriented manifold with orientation chosen counterclockwise.

If a surface is embedded in 3-space, one can choose a continuous unit normal vector n at every point. Then an orientation is defined by the right hand rule with respect to n thought of as the thumb (agudal).

17.3. Cycles and boundaries

The singular homology groups with integer coefficients, $H_k(M;\mathbb{Z})$ for $k = 0, 1, \ldots$ of M are abelian groups which are homotopy invariants of M. Developing the singular homology theory is time-consuming. The case that we will be primarily interested in as far as these notes are concerned, is that of the 1-dimensional homology group:

$$H_1(M;\mathbb{Z}).$$

In this case, the homology groups can be characterized easily without the general machinery of singular simplices.

Let $S^1 \subseteq \mathbb{C}$ be the unit circle, which we think of as a 1-dimensional manifold with an orientation given by the counterclockwise direction.

Definition 17.3.1. A 1-cycle α on a manifold M is an integer linear combination

$$\alpha = \sum_i n_i f_i$$

where $n_i \in \mathbb{Z}$ is called the *multiplicity* (ribui), while each

$$f_i : S^1 \to M$$

is a loop given by a smooth map from the circle to M, and each loop is endowed with the orientation coming from S^1.

Definition 17.3.2. The space of 1-cycles on M is denoted

$$Z_1(M;\mathbb{Z}).$$

Let $(\Sigma_g, \partial\Sigma_g)$ be a surface with boundary $\partial\Sigma_g$, where the genus g is irrelevant for the moment and is only added so as to avoid confusion with the summation symbol $\sum$.

The boundary $\partial\Sigma_g$ is a disjoint union of circles. Now assume the surface Σ_g is oriented.

Proposition 17.3.3. *The orientation of the surface induces an orientation on each boundary circle.*

Thus we obtain an orientation-preserving identification of each boundary component with the standard unit circle $S^1 \subseteq \mathbb{C}$ (with its counterclockwise orientation).

Given a map $\Sigma_g \to M$, its restriction to the boundary therefore produces a 1-cycle

$$\partial\Sigma_g \in Z_1(M;\mathbb{Z}).$$

DEFINITION 17.3.4. The space

$$B_1(M;\mathbb{Z}) \subseteq Z_1(M;\mathbb{Z})$$

of 1-boundaries in M is the space of all cycles

$$\sum_i n_i f_i \in Z_1(M;\mathbb{Z})$$

such that there exists a map of an oriented surface $\Sigma_g \to M$ (for some g) satisfying

$$\partial\Sigma_g = \sum_i n_i f_i.$$

EXAMPLE 17.3.5. Consider the cylinder

$$x^2 + y^2 = 1, \quad 0 \le z \le 1$$

of unit height. The two boundary components correspond to the two circles: the "bottom" circle C_{bottom} defined by $z = 0$, and the "top" circle C_{top} defined by and $z = 1$. Consider the orientation on the cylinder defined by the outward pointing normal vector. It induces the counterclockwise orientation on C_{bottom}, and a clockwise orientation on C_{top}.

Now let C_0 and C_1 be the same circles with the following choice of orientation: we choose a standard counterclockwise parametrisation on both circles, i.e., parametrize them by means of $(\cos\theta, \sin\theta)$. Then the boundary of the cylinder is the *difference* of the two circles: $C_0 - C_1$, or $C_1 - C_0$, depending on the choice of orientation.

EXAMPLE 17.3.6. Cutting up a circle of genus 2 into two once-holed tori shows that the separating curve is a 1-boundary.

THEOREM 17.3.7. *On a closed orientable surface, a separating curve is a boundary, while a non-separating loop is never a boundary.*

17.4. First singular homology group

DEFINITION 17.4.1. The 1-dimensional homology group of M with integer coefficients is the quotient group

$$H_1(M;\mathbb{Z}) = Z_1(M;\mathbb{Z})/B_1(M;\mathbb{Z}).$$

DEFINITION 17.4.2. Given a cycle $C \in Z_1(M;\mathbb{Z})$, its homology class will be denoted $[C] \in H_1(M;\mathbb{Z})$.

EXAMPLE 17.4.3. A non-separating loop on a closed surface represents a non-trivial homology class of the surface.

THEOREM 17.4.4. *The* 1*-dimensional homology group* $H_1(M;\mathbb{Z})$ *is the abelianisation of the fundamental group* $\pi_1(M)$*:*

$$H_1(M;\mathbb{Z}) = (\pi_1 M)^{ab}.$$

Note that a significant difference between the fundamental group and the first homology group is the following. While only based loops participate in the definition of the fundamental group, the definition of $H_1(M;\mathbb{Z})$ involves free (not based) loops.

EXAMPLE 17.4.5. The fundamental groups of the real projective plane $\mathbb{RP}^2$ and the 2-torus $\mathbb{T}^2$ are already abelian. Therefore one obtains

$$H_1(\mathbb{RP}^2;\mathbb{Z}) = \mathbb{Z}/2\mathbb{Z},$$

and

$$H_1(T^2;\mathbb{Z}) = \mathbb{Z}^2.$$

EXAMPLE 17.4.6. The fundamental group of an orientable closed surface Σ_g of genus g is known to be a group on $2g$ generators with a single relation which is a product of g commutators. Therefore one has

$$H_1(\Sigma_g;\mathbb{Z}) = \mathbb{Z}^{2g}.$$

17.5. Stable norm in 1-dimensional homology

Assume the manifold M has a Riemannian metric. Given a smooth loop $f : S^1 \to M$, we can measure its volume (length) with respect to the metric of M. We will denote this length by

$$\mathrm{vol}(f)$$

with a view to higher-dimensional generalisation.

DEFINITION 17.5.1. The volume (length) of a 1-cycle $C = \sum_i n_i f_i$ is defined as

$$\mathrm{vol}(C) = \sum_i |n_i| \,\mathrm{vol}(f_i).$$

DEFINITION 17.5.2. Let $\alpha \in H_1(M;\mathbb{Z})$ be a 1-dimensional homology class. We define the volume of α as the infimum of volumes of representative 1-cycles:

$$\mathrm{vol}(\alpha) = \inf \{\mathrm{vol}(C) \mid C \in \alpha\},$$

where the infimum is over all cycles $C = \sum_i n_i f_i$ representing the class $\alpha \in H_1(M;\mathbb{Z})$.

The following phenomenon occurs for orientable surfaces.

THEOREM 17.5.3. *Let M be an orientable surface, i.e. 2-dimensional manifold. Let $\alpha \in H_1(M;\mathbb{Z})$. For all $j \in \mathbb{N}$, we have*

$$\mathrm{vol}(j\alpha) = j\,\mathrm{vol}(\alpha), \tag{17.5.1}$$

where $j\alpha$ denotes the class $\alpha + \alpha + \ldots + \alpha$, with j summands.

PROOF. To fix ideas, let $j = 2$. By Lemma 17.5.4 below, a minimizing loop C representing a multiple class 2α will necessarily intersect itself in a suitable point p. Then the 1-cycle represented by C can be decomposed into the sum of two 1-cycles (where each can be thought of as a loop based at p). The shorter of the two will then give a minimizing loop in the class α which proves the identity (17.5.1) in this case. The general case follows similarly. □

LEMMA 17.5.4. *A loop going around a cylinder twice necessarily has a point of self-intersection.*

PROOF. We think of the loop as the graph of a 4π-periodic function $f(t)$ (or alternatively a function on $[0, 4\pi]$ with equal values at the endpoints). Consider the difference $g(t) = f(t) - f(t + 2\pi)$. Then g takes both positive and negative values. By the intermediate function theorem, the function g must have a zero t_0. Then $f(t_0) = f(t_0 + 2\pi)$ hence t_0 is a point of self-intersection of the loop. □

17.6. The degree of a map

An example of a degree d map is most easily produced in the case of a circle. A self-map of a circle given by

$$e^{i\theta} \mapsto e^{id\theta}$$

has degree d.

We will discuss the degree in the context of surfaces only.

DEFINITION 17.6.1. The degree

$$d_f$$

of a map f between closed surfaces is the algebraic number of points in the inverse image of a generic point of the target surface.

We can use the 2-dimensional homology groups defined elsewhere in these notes, so as to calculate the degree as follows. Recall that

$$H_2(M;\mathbb{Z}) = \mathbb{Z},$$

where the generator is represented by the identity self-map of the surface.

THEOREM 17.6.2. *A map*

$$f : M_1 \to M_2$$

induces a homomorphism

$$f_* : H_2(M_1; \mathbb{Z}) \to H_2(M_2; \mathbb{Z}),$$

corresponding to multiplication by the degree d_f once the groups are identified with $\mathbb{Z}$:

$$\mathbb{Z} \to \mathbb{Z}, \quad n \mapsto d_f\, n.$$

17.7. Degree of normal map of an embedded surface

THEOREM 17.7.1. *Let $\Sigma \subseteq \mathbb{R}^3$ be an embedded surface. Let p be its genus. Consider the normal map*

$$f_n : \Sigma \to S^2$$

defined by sending each point $x \in \Sigma$ to the normal vector $n = n_x$ at x. Then the degree of the normal map is precisely $1 - p$.

EXAMPLE 17.7.2. For the unit sphere, the normal map is the identity map. The genus is 0, while the degree of the normal map is $1-p = 1$.

EXAMPLE 17.7.3. For the torus, the normal map is harder to visualize. The genus is 1, while the degree of the normal map is 0.

EXAMPLE 17.7.4. For a genus 2 surface embedded in $\mathbb{R}^3$, the degree of the normal map is $1 - 2 = -1$. This means that if the surface is oriented by the outward-pointing normal vector, the normal map is orientation-reversing.

17.8. Euler characteristic of an orientable surface

The Euler characteristic $\chi(M)$ is even for closed orientable surfaces, and the integer $p = p(M) \geq 0$ defined by

$$\chi(M) = 2 - 2p$$

is called the *genus* of M.

EXAMPLE 17.8.1. We have $p(S^2) = 0$, while $p(\mathbb{T}^2) = 1$.

In general, the genus can be understood intuitively as the number of "holes", *i.e.* "handles", in a familiar 3-dimensional picture of a pretzel. We see from formula (17.9.1) that the only compact orientable surface admitting flat metrics is the 2-torus. See [**Ar83**] for a friendly topological introduction to surfaces, and [**Hat02**] for a general definition of the Euler characteristic.

17.9. Gauss–Bonnet theorem

Every embedded closed surface in 3-space admits a continuous choice of a unit normal vector $n = n_x$ at every point x. Note that no such choice is possible for an embedding of the Mobius band, see [**Ar83**] for more details on orientability and embeddings.

Closed embedded surfaces in $\mathbb{R}^3$ are called *orientable*.

Remark 17.9.1. The integrals of type

$$\int_M$$

will be understood in the sense of Theorem 16.9.2, namely using an implied partition subordinate to an atlas, and calculating the integral using coordinates (u^1, u^2) in each chart, so that we can express the metric in terms of metric coefficients

$$g_{ij} = g_{ij}(u^1, u^2)$$

and similarly the Gaussian curvature

$$K = K(u^1, u^2).$$

Theorem 17.9.2 (Gauss–Bonnet theorem). *Every closed surface M satisfies the identity*

$$\int_M K(u^1, u^2)\sqrt{\det(g_{ij})}du^1 du^2 = 2\pi\chi(M), \tag{17.9.1}$$

where K is the Gaussian curvature function on M, whereas $\chi(M)$ is its Euler characteristic.

17.10. Change of metric exploiting Gaussian curvature

We will use the term pseudometric for a quadratic form (or the associated bilinear form), possibly degenerate.

We would like to give an indication of a proof of the Gauss–Bonnet theorem. We will have to avoid discussing some technical points. Consider the normal map

$$F : M \to S^2, \quad x \mapsto n_x.$$

Consider a neigborhood in M where the map F is a homeomorphism (this is not always possible, and is one of the technical points we are avoiding).

Definition 17.10.1. Let g_M the metric of M, and h the standard metric of S^2.

Given a point $x \in M$ in such a neighborhood, we can calculate the curvature $K(x)$. We can then consider a new metric in the conformal class of the metric g_M, defined as follows.

DEFINITION 17.10.2. We define a new pseudometric, denoted $\hat{g}_M$, on M by multiplying by the conformal factor $K(x)$ at the point x. Namely, $\hat{g}_M$ is the pseudometric which at the point x is given by the quadratic form

$$\hat{g}_x = K(x) g_x.$$

If $K \geq 0$ then the length of vectors is multiplied by $\sqrt{K}$.

The key to understanding the proof of the Gauss–Bonnet theorem in the case of embedded surfaces is the following theorem.

THEOREM 17.10.3. *Consider the restriction of the normal map F to a neighborhood as above. We modify the metric on the source M by the conformal factor given by the Gaussian curvature, as above. Then the map*

$$F : (M, \hat{g}_M) \to (S^2, h)$$

preserves areas: the area of the neighborhood in M (with respect to the modified metric) equals to the "area" of its image on the sphere.

17.11. Gauss map

DEFINITION 17.11.1. The Gauss map is the map

$$F : M \to S^2, \quad p \mapsto N_p$$

defined by sending a point p of M the unit normal vector $N = N_p$ thought of as a point of S^2.

The map F sends an infinitesimal parallelogram on the surface, to an infinitesimal parallelogram on the sphere.

We may identify the tangent space to M at p and the tangent space to S^2 at $F(p) \in S^2$. Then the differential of the map F is the Weingarten map

$$W : T_p M \to T_{F(p)} S^2.$$

The element of area KdA of the surface is mapped to the element of area of the sphere. In other words, we modify the element of area by multiplying by the determinant (Jacobian) of the Weingarten map, namely the Gaussian curvature $K(p)$. Hence the image of the area element KdA is precisely area 2-form h on the sphere, as discussed in the previous section.

It remains to be checked that the map has topological degree given by half the Euler characteristic of the surface M, proving the theorem in the case of embedded surfaces. Since degree is invariant under continuous deformations, the result can be checked for a particular standard embedding of a surface of arbitrary genus in $\mathbb{R}^3$.

17.12. An identity

Another way of writing identity (17.5.1) is as follows:

$$\mathrm{vol}(\alpha) = \frac{1}{j}\,\mathrm{vol}(j\alpha).$$

This phenomenon is no longer true for higher-dimensional manifolds. Namely, the volume of a homology class is no longer multiplicative. However, the limit as $j \to \infty$ exists and is called the stable norm.

DEFINITION 17.12.1. Let M be a compact manifold of arbitrary dimension. The *stable norm* $\|\ \|$ of a class $\alpha \in H_1(M;\mathbb{Z})$ is the limit

$$\|\alpha\| = \lim_{j\to\infty} \frac{1}{j}\,\mathrm{vol}(j\alpha). \tag{17.12.1}$$

It is obvious from the definition that one has $\|\alpha\| \le \mathrm{vol}(\alpha)$. However, the inequality may be strict in general. As noted above, for 2-dimensional manifolds we have $\|\alpha\| = \mathrm{vol}(\alpha)$.

PROPOSITION 17.12.2. *The stable norm vanishes for a class of finite order.*

PROOF. If $\alpha \in H_1(M,\mathbb{Z})$ is a class of finite order, one has finitely many possibilities for $\mathrm{vol}(j\alpha)$ as j varies. The factor of $\frac{1}{j}$ in (17.12.1) shows that $\|\alpha\| = 0$. $\square$

Similarly, if two classes differ by a class of finite order, their stable norms coincide. Thus the stable norm passes to the quotient lattice defined below.

DEFINITION 17.12.3. The torsion subgroup of $H_1(M;\mathbb{Z})$ will be denoted $T_1(M) \subset H_1(M;\mathbb{Z})$. The quotient lattice $L_1(M)$ is the lattice

$$L_1(M) = H_1(M;\mathbb{Z})/T_1(M).$$

PROPOSITION 17.12.4. *The lattice $L_1(M)$ is isomorphic to $\mathbb{Z}^{b_1(M)}$, where b_1 is called the first Betti number of M.*

PROOF. This is a general result in the theory of finitely generated abelian groups. $\square$

17.13. Stable systole

DEFINITION 17.13.1. Let M be a manifold endowed with a Riemannian metric, and consider the associated stable norm $\| \ \|$. The stable 1-systole of M, denoted $\mathrm{stsys}_1(M)$, is the least norm of a 1-homology class of infinite order:

$$\begin{aligned}\mathrm{stsys}_1(M) &= \inf\left\{\|\alpha\| \mid \alpha \in H_1(M,\mathbb{Z}) \setminus T_1(M)\right\} \\ &= \lambda_1\left(L_1(M), \| \ \|\right).\end{aligned}$$

EXAMPLE 17.13.2. For an arbitrary metric on the 2-torus $\mathbb{T}^2$, the 1-systole and the stable 1-systole coincide by Theorem 17.5.3:

$$\mathrm{sys}_1(\mathbb{T}^2) = \mathrm{stsys}_1(\mathbb{T}^2),$$

for every metric on $\mathbb{T}^2$.

17.14. Free loops, based loops, and fundamental group

One can refine the notion of simple connectivity by introducing a group, denoted

$$\pi_1(X) = \pi_1(X, x_0),$$

and called the fundamental group of X relative to a fixed "base" point $x_0 \in X$.

DEFINITION 17.14.1. A *based loop* is a loop $\alpha : [0,1] \to X$ satisfying the condition $\alpha(0) = \alpha(1) = x_0$.

In terms of the second item of Definition 17.1.1, we choose a fixed point $s_0 \in S^1$. For example, we can choose $s_0 = e^{i0} = 1$ for the usual unit circle $S^1 \subseteq \mathbb{C}$, and require that $\alpha(s_0) = x_0$. Then the group $\pi_1(X)$ is the quotient of the space of all based loops modulo the equivalence equivalence relation defined by homotopies fixing the basepoint, *cf.* Definition 17.14.2.

The equivalence class of the identity element is precisely the set of contractible loops based at x_0. The equivalence relation can be described as follows for a pair of loops in terms of the second item of Definition 17.1.1.

DEFINITION 17.14.2. Two based loops $\alpha, \beta : S^1 \to X$ are equivalent, or *homotopic*, if there is a continuous map of the cylinder $S^1 \times [c,d] \to X$ whose restriction to $S^1 \times \{c\}$ is α, whose restriction to $S^1 \times \{d\}$ is β, while the map is constant on the segment $\{s_0\} \times [c,d]$ of the cylinder, *i.e.* the basepoint does not move during the homotopy.

Definition 17.14.3. An equivalence class of based loops is called a *based homotopy class*. Removing the basepoint restriction (as well as the constancy condition of Definition 17.14.2), we obtain a larger class called a *free homotopy class (of loops)*.

Definition 17.14.4. *Composition* of two loops is defined most conveniently in terms of item 1 of Definition 17.1.1, by concatenating their domains.

In more detail, the product of a pair of loops, $\alpha : [-1, 0] \to X$ and $\beta : [0, +1] \to X$, is a loop $\alpha.\beta : [-1, 1] \to X$, which coincides with α and β in their domains of definition. The product loop $\alpha.\beta$ is continuous since $\alpha(0) = \beta(0) = x_0$. Then Theorem 17.1.5 can be refined as follows.

17.15. Fundamental groups of surfaces

Theorem 17.15.1. *We have $\pi_1(S^1) = \mathbb{Z}$, while $\pi_1(S^n)$ is the trivial group for all $n \geq 2$.*

Definition 17.15.2. The 2-torus $\mathbb{T}^2$ is defined to be the following Cartesian product: $\mathbb{T}^2 = S^1 \times S^1$, and can thus be realized as a subset

$$\mathbb{T}^2 = S^1 \times S^1 \subseteq \mathbb{R}^2 \times \mathbb{R}^2 = \mathbb{R}^4.$$

We have $\pi_1(\mathbb{T}^2) = \mathbb{Z}^2$. The familiar doughnut picture realizes $\mathbb{T}^2$ as a subset in Euclidean space $\mathbb{R}^3$.

Definition 17.15.3. A 2-dimensional closed Riemannian manifold (*i.e.* surface) M is called *orientable* if it can be realized by a subset of $\mathbb{R}^3$.

Theorem 17.15.4. *The fundamental group of a surface of genus g is isomorphic to a group on $2g$ generators $a_1, b_1, \ldots, a_g, b_g$ satisfying the unique relation*

$$\prod_{i=1}^{g} a_i b_i a_i^{-1} b_i^{-1} = 1.$$

Note that this is not the only possible presentation of the group in terms of a single relation.

CHAPTER 18

Pu's inequality

See `http://u.math.biu.ac.il/~katzmik/egreg826.pdf`

CHAPTER 19

Approach To Loewner via energy-area identity

19.1. An integral-geometric identity

Let $\mathbb{T}^2$ be a torus with an arbitrary metric. Let $\mathbb{T}_0 = \mathbb{R}^2/L$ be the flat torus conformally equivalent to $\mathbb{T}^2$. Let $\ell_0 = \ell_0(x)$ be a simple closed geodesic of $\mathbb{T}_0$. Thus ℓ_0 is the projection of a line $\tilde{\ell}_0 \subseteq \mathbb{R}^2$. Let $\tilde{\ell}_y \subseteq \mathbb{R}^2$ be the line parallel to $\tilde{\ell}_0$ at distance $y > 0$ from ℓ_0 (here we must "choose sides", *e.g.* by orienting $\tilde{\ell}_0$ and requiring $\tilde{\ell}_y$ to lie to the left of $\tilde{\ell}_0$). Denote by $\ell_y \subseteq \mathbb{T}_0$ the closed geodesic loop defined by the image of $\tilde{\ell}_y$. Let $y_0 > 0$ be the *smallest* number such that $\ell_{y_0} = \ell_0$, *i.e.* the lines $\tilde{\ell}_{y_0}$ and $\tilde{\ell}_0$ both project to ℓ_0.

Note that the loops in the family $\{\ell_y\} \subseteq \mathbb{T}^2$ are not necessarily geodesics with respect to the metric of $\mathbb{T}^2$. On the other hand, the family satisfies the following identity.

Lemma 19.1.1 (An elementary integral-geometric identity). *The metric on $\mathbb{T}^2$ satisfies the following identity:*

$$\operatorname{area}(\mathbb{T}^2) = \int_0^{y_0} E(\ell_y)dy, \tag{19.1.1}$$

where E is the energy of a loop with respect to the metric of $\mathbb{T}^2$, see Definition 16.12.2.

Proof. Denote by f^2 the conformal factor of $\mathbb{T}^2$ with respect to the flat metric $\mathbb{T}_0$. Thus the metric on $\mathbb{T}^2$ can be written as

$$f^2(x,y)(dx^2 + dy^2).$$

By Fubini's theorem applied to a rectangle with sides $\operatorname{length}_{\mathbb{T}_0}(\ell_0)$ and y_0, combined with Theorem 16.10.3, we obtain

$$\begin{aligned}\operatorname{area}(\mathbb{T}^2) &= \int_{\mathbb{T}_0} f^2 dxdy \\ &= \int_0^{y_0} \left(\int_{\ell_y} f^2(x,y)dx \right) dy \\ &= \int_0^{y_0} E(\ell_y)dy,\end{aligned}$$

proving the lemma. □

REMARK 19.1.2. The identity (19.1.1) can be thought of as the simplest integral-geometric identity.

19.2. Two proofs of the Loewner inequality

We give a slightly modified version of M. Gromov's proof [**Gro96**], using conformal representation and the Cauchy-Schwarz inequality, of the Loewner inequality (16.2.1) for the 2-torus, see also [**CK03**]. We present the following slight generalisation: there exists a pair of closed geodesics on $(\mathbb{T}^2, g)$, of respective lengths λ_1 and λ_2, such that

$$\lambda_1 \lambda_2 \leq \gamma_2 \operatorname{area}(g), \tag{19.2.1}$$

and whose homotopy classes form a generating set for $\pi_1(\mathbb{T}^2) = \mathbb{Z} \times \mathbb{Z}$.

PROOF. The proof relies on the conformal representation

$$\phi : \mathbb{T}_0 \to (\mathbb{T}^2, g),$$

where $\mathbb{T}_0$ is flat, *cf.* uniformisation theorem 13.12.2. Here the map ϕ may be chosen in such a way that $(\mathbb{T}^2, g)$ and $\mathbb{T}_0$ have the same area. Let f be the conformal factor, so that

$$g = f^2 \left((du^1)^2 + (du^2)^2 \right)$$

as in formula (16.10.1), where $(du^1)^2 + (du^2)^2$ (locally) is the flat metric.

Let ℓ_0 be any closed geodesic in $\mathbb{T}_0$. Let $\{\ell_s\}$ be the family of geodesics parallel to ℓ_0. Parametrize the family $\{\ell_s\}$ by a circle S^1 of length σ, so that

$$\sigma \ell_0 = \operatorname{area}(\mathbb{T}_0).$$

Thus $\mathbb{T}_0 \to S^1$ is a Riemannian submersion. Then

$$\operatorname{area}(\mathbb{T}^2) = \int_{\mathbb{T}_0} f^2.$$

By Fubini's theorem, we have $\operatorname{area}(\mathbb{T}^2) = \int_{S^1} ds \int_{\ell_s} f^2 dt$. Therefore by the Cauchy-Schwarz inequality,

$$\operatorname{area}(\mathbb{T}^2) \geq \int_{S^1} ds \frac{\left(\int_{\ell_s} f dt \right)^2}{\ell_0} = \frac{1}{\ell_0} \int_{S^1} ds \left(\operatorname{length} \phi(\ell_s) \right)^2.$$

Hence there is an s_0 such that $\operatorname{area}(\mathbb{T}^2) \geq \frac{\sigma}{\ell_0} \operatorname{length} \phi(\ell_{s_0})^2$, so that

$$\operatorname{length} \phi(\ell_{s_0}) \leq \ell_0.$$

This reduces the proof to the flat case. Given a lattice in $\mathbb{C}$, we choose a shortest lattice vector λ_1, as well as a shortest one λ_2 not proportional to λ_1. The inequality now follows from Lemma 14.9.1. In the boundary case of equality, one can exploit the equality in the Cauchy-Schwarz inequality to prove that the conformal factor must be constant. □

ALTERNATIVE PROOF. Let ℓ_0 be any simple closed geodesic in $\mathbb{T}_0$. Since the desired inequality (19.2.1) is scale-invariant, we can assume that the loop has unit length:

$$\mathrm{length}_{\mathbb{T}_0}(\ell_0) = 1,$$

and, moreover, that the corresponding covering transformation of the universal cover $\mathbb{C} = \tilde{\mathbb{T}}_0$ is translation by the element $1 \in \mathbb{C}$. We complete 1 to a basis $\{\tau, 1\}$ for the lattice of covering transformations of $\mathbb{T}_0$. Note that the rectangle defined by

$$\{z = x + iy \in \mathbb{C} \mid 0 < x < 1,\ 0 < y < \mathrm{Im}(\tau)]\}$$

is a fundamental domain for $\mathbb{T}_0$, so that $\mathrm{area}(\mathbb{T}_0) = \mathrm{Im}(\tau)$. The maps

$$\ell_y(x) = x + iy,\ x \in [0, 1]$$

parametrize the family of geodesics parallel to ℓ_0 on $\mathbb{T}_0$. Recall that the metric of the torus $\mathbb{T}$ is $f^2(dx^2 + dy^2)$.

LEMMA 19.2.1. *We have the following relation between the length and energy of a loop:*

$$\mathrm{length}(\ell_y)^2 \le E(\ell_y).$$

PROOF. By the Cauchy-Schwarz inequality,

$$\int_0^1 f^2(x, y)dx \ge \left(\int_0^1 f(x, y)dx\right)^2,$$

proving the lemma. □

Now by Lemma 19.1.1 and Lemma 19.2.1, we have

$$\mathrm{area}(\mathbb{T}^2) \ge \int_0^{\mathrm{Im}(\tau)} \big(\mathrm{length}(\ell_y)\big)^2\, dy.$$

Hence there is a y_0 such that

$$\mathrm{area}(\mathbb{T}^2) \ge \mathrm{Im}(\tau)\,\mathrm{length}(\ell_{y_0})^2,$$

so that $\mathrm{length}(\ell_{y_0}) \le 1$. This reduces the proof to the flat case. Given a lattice in $\mathbb{C}$, we choose a shortest lattice vector λ_1, as well as a shortest one λ_2 not proportional to λ_1. The inequality now follows from Lemma 14.9.1. □

19.3. Remark on capacities

Define a conformal invariant called the *capacity* of an annulus as follows. Consider a right circular cylinder

$$\zeta_\kappa = \mathbb{R}/\mathbb{Z} \times [0, \kappa]$$

based on a unit circle $\mathbb{R}/\mathbb{Z}$. Its capacity $C(\zeta_\kappa)$ is defined to be its height, $C(\zeta_\kappa) = \kappa$. Recall that every annular region in the plane is conformally equivalent to such a cylinder, and therefore we have defined a conformal invariant of an arbitrary annular region. Every annular region R satisfies the inequality $\operatorname{area}(R) \geq C(R) \operatorname{sys}_1(R)^2$. Meanwhile, if we cut a flat torus along a shortest loop, we obtain an annular region R with capacity at least $C(R) \geq \gamma_2^{-1} = \frac{\sqrt{3}}{2}$. This provides an alternative proof of the Loewner theorem. In fact, we have the following identity:

$$\operatorname{confsys}_1(g)^2 C(g) = 1, \tag{19.3.1}$$

where confsys is the conformally invariant generalisation of the homology systole, while $C(g)$ is the largest capacity of a cylinder obtained by cutting open the underlying conformal structure on the torus.

Question 19.3.1. Is the Loewner inequality (16.2.1) satisfied by *every* orientable nonsimply connected compact surface? Inspite of its elementary nature, and considerable research devoted to the area, the question is still open. Recently the case of genus 2 was settled as well as genus $g \geq 20$

19.4. A table of optimal systolic ratios of surfaces

Denote by $\mathrm{SR}(M)$ the supremum of the systolic ratios,

$$\mathrm{SR}(M) = \sup_g \frac{\operatorname{sys}_1(g)^2}{\operatorname{area}(g)},$$

ranging over all metrics g on a surface M. The known values of the optimal systolic ratio are tabulated in Figure 19.4.1. It is interesting to note that the optimal ratio for the Klein bottle $\mathbb{RP}^2 \# \mathbb{RP}^2$ is achieved by a singular metric, described in the references listed in the table.

	$\mathrm{SR}(M)$	numerical value	where to find it
$M = \mathbb{RP}^2$	$= \dfrac{\pi}{2}$ [**Pu52**]	≈ 1.5707	site for 88826
infinite $\pi_1(M)$	$< \dfrac{4}{3}$ [**Gro83**]	$< 1.3333\ldots$	(16.1.7)
$M = \mathbb{T}^2$	$= \frac{2}{\sqrt{3}}$ (Loewner)	≈ 1.1547	(16.2.1)
$M = \mathbb{RP}^2 \# \mathbb{RP}^2$	$= \dfrac{\pi}{2^{3/2}}$	≈ 1.1107	[**Bav86, Bav06, Sak88**]
M of genus 2	$> \frac{1}{3}\left(\sqrt{2}+1\right)$	> 0.8047	?
M of genus 3	$\geq \frac{8}{7\sqrt{3}}$	> 0.6598	[**Cal96**]

FIGURE 19.4.1. Values for optimal systolic ratio SR of surface M

CHAPTER 20

A primer on surfaces

In this Chapter, we collect some classical facts on Riemann surfaces. More specifically, we deal with hyperelliptic surfaces, real surfaces, and Katok's optimal bound for the entropy of a surface.

20.1. Hyperelliptic involution

Let M be an orientable closed Riemann surface which is not a sphere. By a Riemann surface, we mean a surface equipped with a fixed conformal structure, *cf.* Definition 16.10.4, while all maps are angle-preserving.

Furthermore, we will assume that the genus is at least 2.

DEFINITION 20.1.1. A *hyperelliptic involution* of a Riemann surface M of genus p is a holomorphic (conformal) map, $J : M \to M$, satisfying $J^2 = 1$, with $2p + 2$ fixed points.

DEFINITION 20.1.2. A surface M admitting a hyperelliptic involution will be called a *hyperelliptic surface*.

REMARK 20.1.3. The involution J can be identified with the non-trivial element in the center of the (finite) automorphism group of M (*cf.* [**FK92**, p. 108]) when it exists, and then such a J is unique, *cf.* [**Mi95**, p.204] (recall that the genus is at least 2).

It is known that the quotient of M by the involution J produces a conformal branched 2-fold covering

$$Q : M \to S^2 \tag{20.1.1}$$

of the sphere S^2.

DEFINITION 20.1.4. The $2p + 2$ fixed points of J are called *Weierstrass points*. Their images in S^2 under the ramified double cover Q of formula (20.1.1) will be referred to as *ramification points*.

A notion of a Weierstrass point exists on any Riemann surface, but will only be used in the present text in the hyperelliptic case.

EXAMPLE 20.1.5. In the case $p = 2$, topologically the situation can be described as follows. A simple way of representing the figure 8 contour in the (x, y) plane is by the reducible curve

$$(((x-1)^2 + y^2) - 1)(((x+1)^2 + y^2) - 1) = 0 \tag{20.1.2}$$

(or, alternatively, by the lemniscate $r^2 = \cos 2\theta$ in polar coordinates, *i.e.* the locus of the equation $(x^2 + y^2)^2 = x^2 - y^2$).

Now think of the figure 8 curve of (20.1.2) as a subset of $\mathbb{R}^3$. The boundary of its tubular neighborhood in $\mathbb{R}^3$ is a genus 2 surface. Rotation by π around the x-axis has six fixed points on the surface, namely, a pair of fixed points near each of the points -2, 0, and $+2$ on the x-axis. The quotient by the rotation can be seen to be homeomorphic to the sphere.

A similar example can be repackaged in a metrically more precise way as follows.

EXAMPLE 20.1.6. We start with a round metric on $\mathbb{RP}^2$. Now attach a small handle. The orientable double cover M of the resulting surface can be thought of as the unit sphere in $\mathbb{R}^3$, with two little handles attached at north and south poles, *i.e.* at the two points where the sphere meets the z-axis. Then one can think of the hyperelliptic involution J as the rotation of M by π around the z-axis. The six fixed points are the six points of intersection of M with the z-axis. Furthermore, there is an orientation reversing involution τ on M, given by the restriction to M of the antipodal map in $\mathbb{R}^3$. The composition $\tau \circ J$ is the reflection fixing the xy-plane, in view of the following matrix identity:

$$\begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \tag{20.1.3}$$

Meanwhile, the induced orientation reversing involution τ_0 on S^2 can just as well be thought of as the reflection in the xy-plane. This is because, at the level of the 2-sphere, it is "the same thing as" the composition $\tau \circ J$. Thus the fixed circle of τ_0 is precisely the equator, *cf.* formula (20.3.3). Then one gets a quotient metric on S^2 which is roughly that of the western hemisphere, with the boundary longitude folded in two, The metric has little bulges along the z-axis at north and south poles, which are leftovers of the small handle.

20.2. Hyperelliptic surfaces

For a treatment of hyperelliptic surfaces, see [**Mi95**, p. 60-61]. By [**Mi95**, Proposition 4.11, p. 92], the affine part of a hyperelliptic surface M is defined by a suitable equation of the form

$$w^2 = f(z) \tag{20.2.1}$$

in $\mathbb{C}^2$, where f is a polynomial. On such an affine part, the map J is given by $J(z,w) = (z,-w)$, while the hyperelliptic quotient map $Q : M \to S^2$ is represented by the projection onto the z-coordinate in $\mathbb{C}^2$.

A slight technical problem here is that the map

$$M \to \mathbb{CP}^2, \tag{20.2.2}$$

whose image is the compactification of the solution set of (20.2.1), is not an embedding. Indeed, there is only one point at infinity, given in homogeneous coordinates by $[0 : w : 0]$. This point is a singularity. A way of desingularizing it using an explicit change of coordinates at infinity is presented in [**Mi95**, p. 60-61]. The resulting smooth surface is unique [**DaS98**, Theorem, p. 100].

REMARK 20.2.1. To explain what happens "geometrically", note that there are two points on our affine surface "above infinity". This means that for a large circle $|z| = r$, there are two circles above it satisfying equation (20.2.1) where f has even degree $2p+2$ (for a Weierstrass point we would only have one circle). To see this, consider $z = re^{ia}$. As the argument a varies from 0 to 2π, the argument of $f(z)$ will change by $(2p+2)2\pi$. Thus, if $(re^{ia}, w(a))$ represents a continuous curve on our surface, then the argument of w changes by $(2p+2)\pi$, and hence we end up where we started, and not at $-w$ (as would be the case were the polynomial of odd degree). Thus there are *two* circles on the surface over the circle $|z| = r$. We conclude that to obtain a smooth compact surface, we will need to add two points at infinity, *cf.* discussion around [**FK92**, formula (7.4.1), p. 102].

Thus, the affine part of M, defined by equation (20.2.1), is a Riemann surface with a pair of punctures p_1 and p_2. A neighborhood of each p_i is conformally equivalent to a punctured disk. By replacing each punctured disk by a full one, we obtain the desired compact Riemann surface M. The point at infinity $[0 : w : 0] \in \mathbb{CP}^2$ is the image of both p_i under the map (20.2.2).

20.3. Ovalless surfaces

Denote by M^ι the fixed point set of an involution ι of a Riemann surface M. Let M be a hyperelliptic surface of even genus p. Let J :

$M \to M$ be the hyperelliptic involution, *cf.* Definition 20.1.1. Let $\tau : M \to M$ be a fixed point free, antiholomorphic involution.

LEMMA 20.3.1. *The involution τ commutes with J and descends to S^2. The induced involution $\tau_0 : S^2 \to S^2$ is an inversion in a circle $C_0 = Q(M^{\tau \circ J})$. The set of ramification points is invariant under the action of τ_0 on S^2.*

PROOF. By the uniqueness of J, *cf.* Remark 20.1.3, we have the commutation relation

$$\tau \circ J = J \circ \tau, \tag{20.3.1}$$

cf. relation (20.1.3). Therefore τ descends to an involution τ_0 on the sphere. There are two possibilities, namely, τ is conjugate, in the group of fractional linear transformations, either to the map $z \mapsto \bar{z}$, or to the map $z \mapsto -\frac{1}{\bar{z}}$. In the latter case, τ is conjugate to the antipodal map of S^2.

In the case of even genus, there is an odd number of Weierstrass points in a hemisphere. Hence the inverse image of a great circle is a connected loop. Thus we get an action of $\mathbb{Z}_2 \times \mathbb{Z}_2$ on a loop, resulting in a contradiction.

In more detail, the set of the $2p+2$ ramification points on S^2 is centrally symmetric. Since there is an odd number, $p+1$, of ramification points in a hemisphere, a generic great circle $A \subseteq S^2$ has the property that its inverse image $Q^{-1}(A) \subseteq M$ is connected. Thus both involutions τ and J, as well as $\tau \circ J$, act fixed point freely on the loop $Q^{-1}(A) \subseteq M$, which is impossible. Therefore τ_0 must fix a point. It follows that τ_0 is an inversion in a circle. □

Suppose a hyperelliptic Riemann surface M admits an antiholomorphic involution τ. In the literature, the components of the fixed point set M^τ of τ are sometimes referred to as "ovals". When τ is fixed point free, we introduce the following terminology.

DEFINITION 20.3.2. A hyperelliptic surface (M, J) of even positive genus $p > 0$ is called *ovalless real* if one of the following equivalent conditions is satisfied:

(1) M admits an imaginary reflection, *i.e.* a fixed point free, antiholomorphic involution τ;
(2) the affine part of M is the locus in $\mathbb{C}^2$ of the equation

$$w^2 = -P(z), \tag{20.3.2}$$

where P is a monic polynomial, of degree $2p+2$, with real coefficients, no real roots, and with distinct roots.

LEMMA 20.3.3. *The two ovalless reality conditions of Definition 20.3.2 are equivalent.*

PROOF. A related result appears in [**GroH81**, p. 170, Proposition 6.1(2)]. To prove the implication (2) $\Longrightarrow$ (1), note that complex conjugation leaves the equation invariant, and therefore it also leaves invariant the locus of (20.3.2). A fixed point must be real, but P is positive hence (20.3.2) has no real solutions. There is no real solution at infinity, either, as there are two points at infinity which are not Weierstrass points, since P is of even degree, as discussed in Remark 20.2.1. The desired imaginary reflection τ switches the two points at infinity, and, on the affine part of the Riemann surface, coincides with complex conjugation $(z, w) \mapsto (\bar{z}, \bar{w})$ in $\mathbb{C}^2$.

To prove the implication (1) $\Longrightarrow$ (2), note that by Lemma 20.3.1, the induced involution τ_0 on $S^2 = M/\,J$ may be thought of as complex conjugation, by choosing the fixed circle of τ_0 to be the circle

$$\mathbb{R} \cup \{\infty\} \subseteq \mathbb{C} \cup \{\infty\} = S^2. \tag{20.3.3}$$

Since the surface is hyperelliptic, it is the smooth completion of the locus in $\mathbb{C}^2$ of *some* equation of the form (20.3.2), *cf.* (20.2.1). Here P is of degree $2p+2$ with distinct roots, but otherwise to be determined. The set of roots of P is the set of (the z-coordinates of) the Weierstrass points. Hence the set of roots must be invariant under τ_0. Thus the roots of the polynomial either come in conjugate pairs, or else are real. Therefore P has real coefficients. Furthermore, the leading cofficient of P may be absorbed into the w-coordinate by extracting a square root. Here we may have to rotate w by i, but at any rate the coefficients of P remain real, and thus P can be assumed monic.

If P had a real root, there would be a ramification point fixed by τ_0. But then the corresponding Weierstrass point must be fixed by τ, as well! This contradicts the fixed point freeness of τ. Thus all roots of P must come in conjugate pairs. $\square$

20.4. Katok's entropy inequality

Let (M, g) be a closed surface with a Riemannian metric. Denote by $(\tilde{M}, \tilde{g})$ the universal Riemannian cover of (M, g). Choose a point $\tilde{x}_0 \in \tilde{M}$.

DEFINITION 20.4.1. The volume entropy (or asymptotic volume) $h(M, g)$ of a surface (M, g) is defined by setting

$$h(M, g) = \lim_{R \to +\infty} \frac{\log\left(\mathrm{vol}_{\tilde{g}}\, B(\tilde{x}_0, R)\right)}{R}, \tag{20.4.1}$$

where $\mathrm{vol}_{\tilde{g}} B(\tilde{x}_0, R)$ is the volume (area) of the ball of radius R centered at $\tilde{x}_0 \in \tilde{M}$.

Since M is compact, the limit in (20.4.1) exists, and does not depend on the point $\tilde{x}_0 \in \tilde{M}$ [**Ma79**]. This asymptotic invariant describes the exponential growth rate of the volume in the universal cover.

Definition 20.4.2. The minimal volume entropy, MinEnt, of M is the infimum of the volume entropy of metrics of unit volume on M, or equivalently

$$\mathrm{MinEnt}(M) = \inf_g h(M, g) \operatorname{vol}(M, g)^{\frac{1}{2}} \tag{20.4.2}$$

where g runs over the space of all metrics on M. For an n-dimensional manifold in place of M, one defines MinEnt similarly, by replacing the exponent of vol by $\frac{1}{n}$.

The classical result of A. Katok [**Kato83**] states that every metric g on a closed surface M with negative Euler characteristic $\chi(M)$ satisfies the optimal inequality

$$h(g)^2 \geq \frac{2\pi|\chi(M)|}{\operatorname{area}(g)}. \tag{20.4.3}$$

Inequality (20.4.3) also holds for hom ent(g) [**Kato83**], as well as the topological entropy, since the volume entropy bounds from below the topological entropy (see [**Ma79**]). We recall the following well-known fact, *cf.* [**KatH95**, Proposition 9.6.6, p. 374].

Lemma 20.4.3. *Let (M, g) be a closed Riemannian manifold. Then,*

$$h(M, g) = \lim_{T \to +\infty} \frac{\log(P'(T))}{T} \tag{20.4.4}$$

where $P'(T)$ is the number of homotopy classes of loops based at some fixed point x_0 which can be represented by loops of length at most T.

Proof. Let $x_0 \in M$ and choose a lift $\tilde{x}_0 \in \tilde{M}$. The group

$$\Gamma := \pi_1(M, x_0)$$

acts on $\tilde{M}$ by isometries. The orbit of $\tilde{x}_0$ under Γ is denoted $\Gamma.\tilde{x}_0$. Consider a fundamental domain Δ for the action of Γ, containing $\tilde{x}_0$. Denote by D the diameter of Δ. The cardinal of $\Gamma.\tilde{x}_0 \cap B(\tilde{x}_0, R)$ is bounded from above by the number of translated fundamental domains $\gamma.\Delta$, where $\gamma \in \Gamma$, contained in $B(\tilde{x}_0, R+D)$. It is also bounded from below by the number of translated fundamental domains $\gamma.\Delta$

contained in $B(\tilde{x}_0, R)$. Therefore, we have

$$\frac{\mathrm{vol}(B(\tilde{x}_0, R))}{\mathrm{vol}(M, g)} \leq \mathrm{card}\,(\Gamma.\tilde{x}_0 \cap B(\tilde{x}_0, R)) \leq \frac{\mathrm{vol}(B(\tilde{x}_0, R+D))}{\mathrm{vol}(M, g)}. \tag{20.4.5}$$

Take the log of these terms and divide by R. The lower bound becomes

$$\begin{aligned} &\frac{1}{R}\log\left(\frac{\mathrm{vol}(B(R))}{\mathrm{vol}(g)}\right) = \\ &\quad = \frac{1}{R}\log\left(\mathrm{vol}(B(R))\right) - \frac{1}{R}\log\left(\mathrm{vol}(g)\right), \end{aligned} \tag{20.4.6}$$

and the upper bound becomes

$$\begin{aligned} &\frac{1}{R}\log\left(\frac{\mathrm{vol}(B(R+D))}{\mathrm{vol}(g)}\right) = \\ &\quad = \frac{R+D}{R}\frac{1}{R+D}\log(\mathrm{vol}(B(R+D))) - \frac{1}{R}\log(\mathrm{vol}(g)). \end{aligned} \tag{20.4.7}$$

Hence both bounds tend to $h(g)$ when R goes to infinity. Therefore,

$$h(g) = \lim_{R\to+\infty}\frac{1}{R}\log\left(\mathrm{card}(\Gamma.\tilde{x}_0 \cap B(\tilde{x}_0, R))\right). \tag{20.4.8}$$

This yields the result since $P'(R) = \mathrm{card}(\Gamma.\tilde{x}_0 \cap B(\tilde{x}_0, R))$. □

Bibliography

[Ab81] Abikoff, W.: The uniformization theorem. *Amer. Math. Monthly* **88** (1981), no. 8, 574–592.

[Ac60] Accola, R.: Differentials and extremal length on Riemann surfaces. *Proc. Nat. Acad. Sci. USA* **46** (1960), 540–543.

[Am04] Ammann, B.: Dirac eigenvalue estimates on two-tori. *J. Geom. Phys.* **51** (2004), no. 3, 372–386.

[Ar83] Armstrong, M.: Basic topology. Corrected reprint of the 1979 original. *Undergraduate Texts in Math.* Springer-Verlag, New York-Berlin, 1983.

[Ar74] Arnold, V.: Mathematical methods of classical mechanics. Translated from the 1974 Russian original by K. Vogtmann and A. Weinstein. Corrected reprint of the second (1989) edition. Graduate Texts in Mathematics, 60. Springer-Verlag, New York, 199?.

[1] Arnol'd, V. I. On the teaching of mathematics.(Russian) Uspekhi Mat. Nauk 53 (1998), no. 1(319), 229–234; translation in Russian Math. Surveys 53 (1998), no. 1, 229–236.

[Ba93] Babenko, I.: Asymptotic invariants of smooth manifolds. *Russian Acad. Sci. Izv. Math.* **41** (1993), 1–38.

[Ba02a] Babenko, I. Forte souplesse intersystolique de variétés fermées et de polyèdres. *Annales de l'Institut Fourier* **52**, 4 (2002), 1259-1284.

[Ba02b] Babenko, I. Loewner's conjecture, Besicovich's example, and relative systolic geometry. [Russian]. *Mat. Sbornik* **193** (2002), 3-16.

[Ba04] Babenko, I.: Géométrie systolique des variétés de groupe fondamental $\mathbb{Z}_2$, *Sémin. Théor. Spectr. Géom. Grenoble* **22** (2004), 25-52.

[BB05] Babenko, I.; Balacheff, F.: Géométrie systolique des sommes connexes et des revêtements cycliques, *Mathematische Annalen* (to appear).

[BaK98] Babenko, I.; Katz, M.: Systolic freedom of orientable manifolds, *Annales Scientifiques de l'E.N.S. (Paris)* **3**1 (1998), 787-809.

[Bana93] Banaszczyk, W. New bounds in some transference theorems in the geometry of numbers. *Math. Ann.* **296** (1993), 625–635.

[Bana96] Banaszczyk, W. Inequalities for convex bodies and polar reciprocal lattices in $\mathbb{R}^n$ II: Application of K-convexity. *Discrete Comput. Geom.* **16** (1996), 305–311.

[BCIK04] Bangert, V; Croke, C.; Ivanov, S.; Katz, M. Boundary case of equality in optimal Loewner-type inequalities. *Trans. Amer. Math. Soc.* **359** (2007), no. 1, 1–17. See `https://arxiv.org/abs/math/0406008`

[BCIK05] Bangert, V; Croke, C.; Ivanov, S.; Katz, M.: Filling area conjecture and ovalless real hyperelliptic surfaces, *Geometric and Functional Analysis (GAFA)* **15** (2005) no. 3. See `arXiv:math.DG/0405583`

[BK03] Bangert, V.; Katz, M.: Stable systolic inequalities and cohomology products, *Comm. Pure Appl. Math.* **56** (2003), 979–997. Available at `arXiv:math.DG/0204181`

[BK04] Bangert, V; Katz, M.: An optimal Loewner-type systolic inequality and harmonic one-forms of constant norm. *Comm. Anal. Geom.* **12** (2004), number 3, 703-732. See `arXiv:math.DG/0304494`

[Bar57] Barnes, E. S.: On a theorem of Voronoi. *Proc. Cambridge Philos. Soc.* **53** (1957), 537–539.

[Bav86] Bavard, C.: Inégalité isosystolique pour la bouteille de Klein. *Math. Ann.* **274** (1986), no. 3, 439–441.

[Bav88] Bavard, C.: Inégalités isosystoliques conformes pour la bouteille de Klein. *Geom. Dedicata* **27** (1988), 349–355.

[Bav92] Bavard, C.: La systole des surfaces hyperelliptiques, *Prepubl. Ec. Norm. Sup. Lyon* **71** (1992).

[Bav06] Bavard, C.: Une remarque sur la géométrie systolique de la bouteille de Klein. *Arch. Math. (Basel)* **87** (2006), no. 1, 72–74.

[BeM] Bergé, A.-M.; Martinet, J. Sur un problème de dualité lié aux sphères en géométrie des nombres. *J. Number Theory* **32** (1989), 14–42.

[Ber72] Berger, M. A l'ombre de Loewner. *Ann. Scient. Ec. Norm. Sup. (Sér. 4)* **5** (1972), 241–260.

[Ber80] Berger, M.: Une borne inférieure pour le volume d'une variété riemannienne en fonction du rayon d'injectivité. *Ann. Inst. Fourier* **30** (1980), 259–265.

[Ber93] Berger, M. Systoles et applications selon Gromov. Séminaire N. Bourbaki, exposé 771, *Astérisque* **216** (1993), 279–310.

[Ber76] Berstein, I.: On the Lusternik–Schnirelmann category of Grassmannians, *Proc. Camb. Phil. Soc.* **79** (1976), 129-134.

[Bes87] Besse, A.: Einstein manifolds. *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*, **10**. Springer-Verlag, Berlin, 1987.

[BesCG95] Besson, G.; Courtois, G.;, Gallot, S.: Volume et entropie minimale des espaces localement symétriques. *Invent. Math.* **103** (1991), 417–445.

[BesCG95] Besson, G.; Courtois, G.;, Gallot, S.: Hyperbolic manifolds, amalgamated products and critical exponents. *C. R. Math. Acad. Sci. Paris* **336** (2003), no. 3, 257–261.

[Bl61] Blatter, C. Über Extremallängen auf geschlossenen Flächen. *Comment. Math. Helv.* **35** (1961), 153–168.

[Bl61b] Blatter, C.: Zur Riemannschen Geometrie im Grossen auf dem Möbiusband. *Compositio Math.* **15** (1961), 88–107.

[Borovik & Katz 2012] Borovik, A., Katz, M. "Who gave you the Cauchy–Weierstrass tale? The dual history of rigorous calculus." *Foundations of Science* **17**, no. 3, 245-276.
See `http://dx.doi.org/10.1007/s10699-011-9235-x`

[Bo94] Borzellino, J. E.: Pinching theorems for teardrops and footballs of revolution. *Bull. Austral. Math. Soc.* **49** (1994), no. 3, 353–364.

[BG76] Bousfield, A.; Gugenheim, V.: On PL de Rham theory and rational homotopy type. *Mem. Amer. Math. Soc.* **8**, 179, Providence, R.I., 1976.

[BI94] Burago, D.; Ivanov, S.: Riemannian tori without conjugate points are flat. *Geom. Funct. Anal.* **4** (1994), no. 3, 259–269.

[BI95] Burago, D.; Ivanov, S.: On asymptotic volume of tori. *Geom. Funct. Anal.* **5** (1995), no. 5, 800–808.

[BuraZ80] Burago, Y.; Zalgaller, V.: Geometric inequalities. *Grundlehren der Mathematischen Wissenschaften* **285**. Springer Series in Soviet Mathematics. Springer-Verlag, Berlin, 1988 (original Russian edition: 1980).

[BS94] Buser, P.; Sarnak, P.: On the period matrix of a Riemann surface of large genus. With an appendix by J. H. Conway and N. J. A. Sloane. *Invent. Math.* **117** (1994), no. 1, 27–56.

[Cal96] Calabi, E.: Extremal isosystolic metrics for compact surfaces, Actes de la table ronde de geometrie differentielle, *Sem. Congr.* **1** (1996), Soc. Math. France, 167–204.

[Ca76] Manfredo do Carmo, Differential geometry of curves and surfaces. Prentice-Hall, 1976.

[Car92] do Carmo, M.: Riemannian geometry. Translated from the second Portuguese edition by Francis Flaherty. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992.

[Ca71] Cassels, J. W. S.: An introduction to the geometry of numbers. Second printing. *Grundlehren der mathematischen Wissenschaften*, **9**9. Springer-Verlag, Berlin-Heidelberg-New York, 1971.

[Cauchy 1826] Cauchy, A. L. Leçns sur les applications du calcul infinitésimal à la géométrie. Paris: Imprimérie Royale, 1826.

[Ch93] Chavel, I.: Riemannian geometry—a modern introduction. *Cambridge Tracts in Mathematics*, **1**08. Cambridge University Press, Cambridge, 1993.

[CoS94] Conway, J.; Sloane, N.: On lattices equivalent to their duals. J. Number Theory **48** (1994), no. 3, 373–382.

[ConS99] Conway, J. H.; Sloane, N. J. A.: Sphere packings, lattices and groups. Third edition. Springer, 1999.

[CoT03] Cornea, O.; Lupton, G.; Oprea, J.; Tanré, D.: Lusternik-Schnirelmann category. *Mathematical Surveys and Monographs*, **103**. American Mathematical Society, Providence, RI, 2003.

[Cr80] Croke, C.: Some isoperimetric inequalities and eigenvalue estimates. *Ann. Sci. École Norm. Sup.* **13** (1980), 519–535.

[Cr88] Croke, C.: An isoembolic pinching theorem. *Invent. Math.* **92** (1988), 385–387.

[CK03] Croke, C.; Katz, M.: Universal volume bounds in Riemannian manifolds, *Surveys in Differential Geometry* **V**III, Lectures on Geometry and Topology held in honor of Calabi, Lawson, Siu, and Uhlenbeck at Harvard University, May 3- 5, 2002, edited by S.T. Yau (Somerville, MA: International Press, 2003.) pp. 109 - 137. See `arXiv:math.DG/0302248`

[DaS98] Danilov, V. I.; Shokurov, V. V.: Algebraic curves, algebraic manifolds and schemes. Translated from the 1988 Russian original by D. Coray and V. N. Shokurov. Reprint of the original English edition from the series Encyclopaedia of Mathematical Sciences [Algebraic geometry. I, Encyclopaedia Math. Sci., 23, Springer, Berlin, 1994]. Springer-Verlag, Berlin, 1998.

[Do79] Dombrowski, P.: 150 years after Gauss' "Disquisitiones generales circa superficies curvas". With the original text of Gauss. *Astérisque*, **62**. Société Mathématique de France, Paris, 1979.

[DR03] Dranishnikov, A.; Rudyak, Y.: Examples of non-formal closed $(k-1)$-connected manifolds of dimensions $\geq 4k-1$, *Proc. Amer. Math. Soc.* **133** (2005), no. 5, 1557-1561. See `arXiv:math.AT/0306299`.

[FK92] Farkas, H. M.; Kra, I.: Riemann surfaces. Second edition. Graduate Texts in Mathematics, 71. Springer-Verlag, New York, 1992.

[Fe69] Federer, H.: Geometric Measure Theory. *Grundlehren der mathematischen Wissenschaften*, **1**53. Springer-Verlag, Berlin, 1969.

[Fe74] Federer, H. Real flat chains, cochains, and variational problems. *Indiana Univ. Math. J.* **24** (1974), 351–407.

[FHL98] Felix, Y.; Halperin, S.; Lemaire, J.-M.: The rational LS category of products and of Poincaré duality complexes. *Topology* **37** (1998), no. 4, 749–756.

[Fer77] Ferrand, J.: Sur la regularite des applications conformes, *C.R. Acad.Sci. Paris Ser.A-B* **2**84 (1977), no.1, A77-A79.

[Fr99] Freedman, M.: $\mathbb{Z}_2$-Systolic-Freedom. *Geometry and Topology Monographs* **2**, Proceedings of the Kirbyfest (J. Hass, M. Scharlemann eds.), 113–123, Geometry & Topology, Coventry, 1999.

[Fu16] Funk, P.: Uber eine geometrische Anwendung der Abelschen Integralgleichung. *Math. Ann.* **7**7 (1916), 129-135.

[GaHL04] Gallot, S.; Hulin, D.; Lafontaine, J.: Riemannian geometry. Third edition. Universitext. Springer-Verlag, Berlin, 2004.

[Ga71] Ganea, T.: Some problems on numerical homotopy invariants. Symposium on Algebraic Topology (Battelle Seattle Res. Center, Seattle Wash., 1971), pp. 23–30. *Lecture Notes in Math.,* **249**, Springer, Berlin, 1971.

[2] Giorgi, Giorgio. Various Proofs of the Sylvester Criterion for Quadratic Forms. *Journal of Mathematics Research* **9** (December 2017), no. 6, 55–66.

[3] Giusti, Enrico. *Bonaventura Cavalieri and the theory of indivisibles.* Bologna, Edizioni Cremonese, 1980.

[Goldman 2005] Goldman, Ron. Curvature formulas for implicit curves and surfaces. *Comput. Aided Geom. Design* **22** (2005), no. 7, 632–658. See http://u.math.biu.ac.il/ katzmik/goldman05.pdf

[GG92] Gómez-Larrañaga, J.; González-Acuña, F.: Lusternik-Schnirel'mann category of 3-manifolds. *Topology* **31** (1992), no. 4, 791–800.

[Gre67] Greenberg, M. Lectures on Algebraic Topology. W.A. Benjamin, Inc., New York, Amsterdam, 1967.

[GHV76] Greub, W.; Halperin, S.; Vanstone, R.: Connections, curvature, and cohomology. Vol. III: Cohomology of principal bundles and homogeneous spaces, *Pure and Applied Math.*, **47**, Academic Press, New York-London, 1976.

[GriH78] Griffiths, P.; Harris, J.: Principles of algebraic geometry. Pure and Applied Mathematics. Wiley-Interscience [John Wiley & Sons], New York, 1978 [1994].

[Gro81a] Gromov, M.: Structures métriques pour les variétés riemanniennes. (French) [Metric structures for Riemann manifolds] Edited by J. Lafontaine and P. Pansu. *Textes Mathématiques*, **1**. CEDIC, Paris, 1981.

[Gro81b] Gromov, M.: Volume and bounded cohomology. *Publ. IHES.* **56** (1981), 213–307.

[Gro83] Gromov, M.: Filling Riemannian manifolds, *J. Diff. Geom.* **18** (1983), 1-147.

[Gro96] Gromov, M.: Systoles and intersystolic inequalities, Actes de la Table Ronde de Géométrie Différentielle (Luminy, 1992), 291–362, *Sémin. Congr.*, **1**, Soc. Math. France, Paris, 1996. `www.emis.de/journals/SC/1996/1/ps/smf_sem-cong_1_291-362.ps.gz`

[Gro99] Gromov, M.: Metric structures for Riemannian and non-Riemannian spaces, *Progr. in Mathematics*, **152**, Birkhäuser, Boston, 1999.

[GroH81] Gross, B. H.; Harris, J.: Real algebraic curves. *Ann. Sci. Ecole Norm. Sup.* (4) **14** (1981), no. 2, 157–182.

[GruL87] Gruber, P.; Lekkerkerker, C.: *Geometry of numbers*. Second edition. North-Holland Mathematical Library, **37**. North-Holland Publishing Co., Amsterdam, 1987.

[Hat02] Hatcher, A.: Algebraic topology. Cambridge University Press, Cambridge, 2002.

[He86] Hebda, J. The collars of a Riemannian manifold and stable isosystolic inequalities. *Pacific J. Math.* **121** (1986), 339-356.

[Hel99] Helgason, S.: The Radon transform. 2nd edition. Progress in Mathematics, 5. Birkhäuser, Boston, Mass., 1999.

[Hem76] Hempel, J.: 3-Manifolds. *Ann. of Math. Studies* **8**6, Princeton Univ. Press, Princeton, New Jersey 1976.

[Hil03] Hillman, J. A.: Tits alternatives and low dimensional topology. *J. Math. Soc. Japan* **55** (2003), no. 2, 365–383.

[Hi04] Hillman, J. A.: An indecomposable PD_3-complex: II, *Algebraic and Geometric Topology* **4** (2004), 1103-1109.

[Iv02] Ivanov, S.: On two-dimensional minimal fillings, *St. Petersbg. Math. J.* **13** (2002), no. 1, 17–25.

[IK04] Ivanov, S.; Katz, M.: Generalized degree and optimal Loewner-type inequalities, *Israel J. Math.* **141** (2004), 221-233. `arXiv:math.DG/0405019`

[Iw02] Iwase, N.: A_∞-method in Lusternik-Schnirelmann category. *Topology* **41** (2002), no. 4, 695–723.

[Jo48] John, F. Extremum problems with inequalities as subsidiary conditions. Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948, 187–204. Interscience Publishers, New York, N. Y., 1948.

[Kato83] Katok, A.: Entropy and closed geodesics, *Ergo. Th. Dynam. Sys.* **2** (1983), 339–365.

[Kato92] Katok, S.: Fuchsian groups. Chicago Lectures in Mathematics. University of Chicago Press, Chicago, IL, 1992.

[KatH95] Katok, A.; Hasselblatt, B.: Introduction to the modern theory of dynamical systems. With a supplementary chapter by Katok and L. Mendoza. *Encyclopedia of Mathematics and its Applications*, **54**. Cambridge University Press, Cambridge, 1995.

[Ka83] Katz, M.: The filling radius of two-point homogeneous spaces, *J. Diff. Geom.* **18** (1983), 505-511.

[Ka88] Katz, M.: The first diameter of 3-manifolds of positive scalar curvature. *Proc. Amer. Math. Soc.* **104** (1988), no. 2, 591–595.

[Ka02] Katz, M. Local calibration of mass and systolic geometry. *Geom. Funct. Anal. (GAFA)* **12**, issue 3 (2002), 598-621. See `arXiv:math.DG/0204182`

[Ka03] Katz, M.: Four-manifold systoles and surjectivity of period map, *Comment. Math. Helv.* **78** (2003), 772-786. See `arXiv:math.DG/0302306`

[Ka05] Katz, M.: Systolic inequalities and Massey products in simply-connected manifolds, *Geometriae Dedicata*, to appear.

[Ka07] Katz, M.: Systolic geometry and topology. With an appendix by Jake P. Solomon. *Mathematical Surveys and Monographs*, **137**. American Mathematical Society, Providence, RI, 2007.

[4] Katz, M.; Leichtnam, E. "Commuting and noncommuting infinitesimals." *American Mathematical Monthly* **120** (2013), no. 7, 631–641. See `http://dx.doi.org/10.4169/amer.math.monthly.120.07.631` and `https://arxiv.org/abs/1304.0583`

[KL04] Katz, M.; Lescop, C.: Filling area conjecture, optimal systolic inequalities, and the fiber class in abelian covers. Proceedings of conference and workshop in memory of R. Brooks, held at the Technion, *Israel Mathematical Conference Proceedings (IMCP), Contemporary Math.*, Amer. Math. Soc., Providence, R.I. (to appear). See `arXiv:math.DG/0412011`

[KR04] Katz, M.; Rudyak, Y.: Lusternik–Schnirelmann category and systolic category of low dimensional manifolds. *Communications on Pure and Applied Mathematics*, to appear. See `arXiv:math.DG/0410456`

[KR05] Katz, M.; Rudyak, Y.: Bounding volume by systoles of 3-manifolds. See `arXiv:math.DG/0504008`

[KatzRS06] Katz, M.; Rudyak, Y.; Sabourau, S.: Systoles of 2-complexes, Reeb graph, and Grushko decomposition. *International Math. Research Notices* 2006 (2006). Art. ID 54936, pp. 1–30. See `arXiv:math.DG/0602009`

[KS04] Katz, M.; Sabourau, S.: Hyperelliptic surfaces are Loewner, *Proc. Amer. Math. Soc.*, to appear. See `arXiv:math.DG/0407009`

[KS05] Katz, M.; Sabourau, S.: Entropy of systolically extremal surfaces and asymptotic bounds, *Ergodic Theory and Dynamical Systems*, **25** (2005). See `arXiv:math.DG/0410312`

[KS06] Katz, M.; Sabourau, S.: An optimal systolic inequality for CAT(0) metrics in genus two, preprint. See `arXiv:math.DG/0501017`

[KSV05] Katz, M.; Schaps, M.; Vishne, U.: Logarithmic growth of systole of arithmetic Riemann surfaces along congruence subgroups. See `arXiv:math.DG/0505007`

[Ke74] Keisler, H. J.: Elementary calculus: An infinitesimal approach. Available online at http://www.math.wisc.edu/∼keisler/calc.html

[Kl82] Klein, F.: Über Riemanns Theorie der algebraischen Funktionen und ihrer Integrale. – Eine Ergänzung der gewöhnlichen Darstellungen. Leipzig: B.G. Teubner 1882.

[Ko87] Kodani, S.: On two-dimensional isosystolic inequalities, *Kodai Math. J.* **10** (1987), no. 3, 314–327.

[Kon03] Kong, J.: Seshadri constants on Jacobian of curves. *Trans. Amer. Math. Soc.* **355** (2003), no. 8, 3175–3180.

[Ku78] Kung, H. T.; Leiserson, C. E.: Systolic arrays (for VLSI). Sparse Matrix Proceedings 1978 (Sympos. Sparse Matrix Comput., Knoxville, Tenn., 1978), pp. 256–282, SIAM, Philadelphia, Pa., 1979.

[LaLS90] Lagarias, J.C.; Lenstra, H.W., Jr.; Schnorr, C.P.: Bounds for Korkin-Zolotarev reduced bases and successive minima of a lattice and its reciprocal lattice. *Combinatorica* **10** (1990), 343–358.

[La75] Lawson, H.B. The stable homology of a flat torus. *Math. Scand.* **36** (1975), 49–73.

[Leib] Leibniz, G. W.: La naissance du calcul différentiel. (French) [The birth of differential calculus] 26 articles des Acta Eruditorum. [26 articles of the Acta Eruditorum] Translated from the Latin and with an introduction and notes by Marc Parmentier. With a preface by Michel Serres. Mathesis. Librairie Philosophique J. Vrin, Paris, 1989.

[Li69] Lichnerowicz, A.: Applications harmoniques dans un tore. *C.R. Acad. Sci., Sér. A* **2**69, 912–916 (1969).

[Lo71] Loewner, C.: Theory of continuous groups. Notes by H. Flanders and M. Protter. *Mathematicians of Our Time* **1**, The MIT Press, Cambridge, Mass.-London, 1971.

[Ma69] Macbeath, A. M.: Generators of the linear fractional groups. 1969 Number Theory (Proc. Sympos. Pure Math., Vol. XII, Houston, Tex., 1967) pp. 14–32 Amer. Math. Soc., Providence, R.I.

[Mi95] Miranda, R.: Algebraic curves and Riemann surfaces. *Graduate Studies in Mathematics*, **5**. American Mathematical Society, Providence, RI, 1995.

[Ma79] Manning, A.: Topological entropy for geodesic flows, *Ann. of Math.* (2) **110** (1979), 567–573.

[Ma69] May, J.: Matric Massey products. *J. Algebra* **12** (1969) 533–568.

[MP77] Millman, R.; Parker, G. Elements of differential geometry. Prentice-Hall Inc., Englewood Cliffs, N. J., 1977.

[MS86] Milman, V. D.; Schechtman, G.: Asymptotic theory of finite-dimensional normed spaces. With an appendix by M. Gromov. *Lecture Notes in Mathematics*, **1200**. Springer-Verlag, Berlin, 1986.

[MilH73] Milnor, J.; Husemoller, D.: Symmetric bilinear forms. Springer, 1973.

[MiTW73] Misner, Charles W.; Thorne, Kip S.; Wheeler, John Archibald: Gravitation. W. H. Freeman and Co., San Francisco, Calif., 1973.

[Mo86] Moser, J. Minimal solutions of variational problems on a torus. *Ann. Inst. H. Poincaré-Analyse non linéaire* **3** (1986), 229–272.

[MS92] Moser, J.; Struwe, M. On a Liouville-type theorem for linear and nonlinear elliptic differential equations on a torus. *Bol. Soc. Brasil. Mat.* **23** (1992), 1–20.

[NR02] Nabutovsky, A.; Rotman, R.: The length of the shortest closed geodesic on a 2-dimansional sphere, *Int. Math. Res. Not.* **2002:23** (2002), 1211-1222.

[OR01] Oprea, J.; Rudyak, Y.: Detecting elements and Lusternik-Schnirelmann category of 3-manifolds. Lusternik-Schnirelmann category and related topics (South Hadley, MA, 2001), 181–191, *Contemp. Math.*, **316**, Amer. Math. Soc., Providence, RI, 2002. `arXiv:math.AT/0111181`

[Pa04] Parlier, H.: Fixed-point free involutions on Riemann surfaces. See `arXiv:math.DG/0504109`

[PP03] Paternain, G. P.; Petean, J.: Minimal entropy and collapsing with curvature bounded from below. *Invent. Math.* **151** (2003), no. 2, 415–450.

[Pu52] Pu, P.M.: Some inequalities in certain nonorientable Riemannian manifolds, *Pacific J. Math.* **2** (1952), 55–71.

[Ri1854] B. Riemann, Hypothèses qui servent de fondement à la géométrie, in Oeuvres mathématiques de Riemann. J. Gabay, 1990.

[RT90] Rosenbloom, M.; Tsfasman, M.: Multiplicative lattices in global fields. *Invent. Math.* **101** (1990), no. 3, 687–696.

[Ru87] Rudin, W.: Real and complex analysis. Third edition. McGraw-Hill Book Co., New York, 1987.

[Ru99] Rudyak, Y.: On category weight and its applications. *Topology* **38**, 1, (1999), 37–55.

[RT00] Rudyak, Y.; Tralle, A.: On Thom spaces, Massey products, and nonformal symplectic manifolds. *Internat. Math. Res. Notices* **10** (2000), 495–513.

[Sa04] Sabourau, S.: Systoles des surfaces plates singulières de genre deux, *Math. Zeitschrift* **247** (2004), no. 4, 693–709.

[Sa05] Sabourau, S.: Entropy and systoles on surfaces, preprint.

[Sa06] Sabourau, S.: Systolic volume and minimal entropy of aspherical manifolds, preprint.

[Sachs] Sachs, Rainer Kurt; Wu, Hung Hsi. General relativity for mathematicians. Graduate Texts in Mathematics, Vol. 48. Springer-Verlag, New York-Heidelberg, 1977.

[Sak88] Sakai, T.: A proof of the isosystolic inequality for the Klein bottle. *Proc. Amer. Math. Soc.* **104** (1988), no. 2, 589–590.

[SS03] Schmidt, Thomas A.; Smith, Katherine M.: Galois orbits of principal congruence Hecke curves. J. London Math. Soc. (2) 67 (2003), no. 3, 673–685.

[Sen91a] Senn, W. Strikte Konvexität für Variationsprobleme auf dem n-dimensionalen Torus. *manuscripta math.* **71** (1991), 45–65.

[Sen91b] Senn, W. Über Mosers regularisiertes Variationsproblem für minimale Blätterungen des n-dimensionalen Torus. *J. Appl. Math. Physics (ZAMP)* **42** (1991), 527–546.

[Ser73] Serre, J.-P.: A course in arithmetic. Translated from the French. Graduate Texts in Mathematics, No. 7. Springer-Verlag, New York-Heidelberg, 1973.

[5] Sherry, D.; Katz, M. Infinitesimals, imaginaries, ideals, and fictions. *Studia Leibnitiana* **44** (2012), no. 2, 166–192. `http://www.jstor.org/stable/43695539`, `https://arxiv.org/abs/1304.2137` (Article was published in 2014 even though the journal issue lists the year as 2012)

[Sik05] Sikorav, J.: Bounds on primitives of differential forms and cofilling inequalities. See `arXiv:math.DG/0501089`

[Sin78] Singhof, W.: Generalized higher order cohomology operations induced by the diagonal mapping. *Math. Z.* **162**, 1, (1978), 7–26.

[Sm62] Smale, S.: On the structure of 5-manifolds. *Ann. of Math.* (2) **75** (1962), 38–46.

[Sp79] Spivak, M.: A comprehensive introduction to differential geometry. Vol. 2, chapter 4. Third edition. Publish or Perish, Inc., Wilmington, Del., 1999.

[6] Stillwell, John. *Mathematics and its history. Second edition.* Undergraduate Texts in Mathematics. Springer-Verlag, New York, 2002.

[St00] Streit, M.: Field of definition and Galois orbits for the Macbeath-Hurwitz curves. Arch. Math. (Basel) 74 (2000), no. 5, 342–349.

[Tr93] Tralle, A.: On compact homogeneous spaces with nonvanishing Massey products, in *Differential geometry and its applications (Opava, 1992)*, pp. 47–50, Math. Publ. **1**, Silesian Univ. Opava, 1993.

[Tu47] Tutte, W.: A family of cubical graphs. *Proc. Cambridge Philos. Soc.* **43** (1947), 459–474.

[7] Valenti, R.; Sebe, N.; Gevers, T. Combining head pose and eye location information for gaze estimation. IEEE Trans. Image Process. 21 (2012), no. 2, 802–815.

[8] Valenti R.; Gevers T. Accurate Eye Center Location and Tracking Using Isophote Curvature. 2014.

[Wa83] Warner, F.W.: Foundations of Differentiable Manifolds and Lie Groups. *Graduate Texts in Mathematics*, **9**4. Springer-Verlag, New York-Heidelberg-Berlin, 1983.

[We55] Weatherburn, C. E.: Differential geometry of three dimensions. Vol. 1. Cambridge University Press, 1955.

[50] White, B.: Regularity of area-minimizing hypersurfaces at boundaries with multiplicity. Seminar on minimal submanifolds (E. Bombieri ed.), 293-301. *Annals of Mathematics Studies* **103**. Princeton University Press, Princeton N.J., 1983.

[Wr74] Wright, Alden H.: Monotone mappings and degree one mappings between PL manifolds. Geometric topology (Proc. Conf., Park City, Utah, 1974), pp. 441–459. Lecture Notes in Math., Vol. 438, Springer, Berlin, 1975.

[Ya88] Yamaguchi, T.: Homotopy type finiteness theorems for certain precompact families of Riemannian manifolds, *Proc. Amer. Math. Soc.* **102** (1988), 660–666.