Geometry of Surfaces

G. B. Segal

Mathematical Institute, Oxford

November 1986

Re-printed August 1995

Contents

Introduction

These notes are slightly different in spirit from most parts of undergraduate pure mathematics. Surfaces are things which everyone can see, and the questions we ask about them are very natural and interesting ones, which - roughly at least - are easily explained to a layman. From the point of view of the foundations of mathematics, however, a smooth surface is a concept of a much higher order of complexity and sophistication than, say, a group or a ring. A geometer is no more concerned with this conceptual sophistication than is a cook with biochemistry. He cannot completely ignore the foundations of the subject without becoming hopelessly imprecise and inaccurate; but if he is to do geometry rather than something else he has to keep the foundations firmly in the background. In a course of the length of this one it would be very easy to devote the whole space to the careful definition of smooth surfaces and smooth maps, and not get to geometry at all. To avoid that I have decided resolutely not to follow the prevailing style of undergraduate exposition, which amounts to spelling out explicitly every idea that is intended to pass through the reader's mind. Although I realize that many students find such a style reassuring I am not convinced that it is healthy, or even that it serves the purpose of making the subject clear and "easy"; I believe that questions of light and shade, and perspective, are essential for real understanding. Thus in these notes I never discuss such points as whether the composite of smooth maps is smooth, and I have put the inverse

and implicit function theorems at the end in an appendix to emphasize that they are not themselves geometry. My feeling is that if a reader is perceptive enough to be disturbed when I define a smooth function on an open interval and later speak of smooth functions on closed intervals without commenting on one-sided derivatives, then he will be able to provide his own remedy; and if he does not notice the point I think it is best to economize his powers by not directing them to diversionary issues.

Another feature of my style which some may not like is that I have frequently mentioned results which are not proved in the course but which I think readers should be aware of: the fact, for example, that every surface can be covered by conformal charts.

The essential part of these notes consists of the local differential geometry of surfaces. Nevertheless I have devoted the first four sections entirely to topology. A number of motives influenced me to do this. One was to show the interplay between local and global properties, which I regard as the most interesting aspect of differential geometry, and which is beautifully exemplified in the Gauss-Bonnet theorem relating the topology of a surface to its curvature. But a more basic motive was just to emphasize that a topological space is an extemely natural concept. I feel sure that the idea of a topology on a set comes as readily to the mind as the idea of a set itself. Anyone who is happy with the set of all colours - I mean colours as one finds them on colour charts in paint shops - will know what

is meant by saying that a colour is changing continuously or discontinuously. It has been one of the triumphs of mathematical formalization to see that the intuitively clear but nevertheless elusive idea of a topology on a set amounts to the knowledge of when a subset is a "neighbourhood" of one of its points, or - less illuminatingly still - the knowledge of which subsets are "open"; but experience has shown me that the formal definition does the opposite of giving undergraduates the right idea. I hope that thinking explicitly about the topology of surfaces will do something to redress the balance.

Considerations of the same kind led me to spend some time discussing "abstract" surfaces - those which do not arise as subsets of $\mathbb{R}^3$. The fact that abstractly defined sets often have a significant geometry is, I believe, one of the most valuable ideas that mathematics has to offer; and at the same time it seems one of the hardest to make clear. (Its importance has been shown most strikingly in recent particle physics.)

In short, readers of these notes who want to do the minimum for the sake of examinations can pass fleetingly over §§1-5, and can ignore "abstract" surfaces and everything concerned with complex numbers completely. The contents of these sections are hardly referred to in the sequel. I hope, all the same, that many readers will find them interesting and profitable.

Finally, what is the position of the theory of surfaces in present-day mathematics? Most geometrical research

nowadays is concerned with manifolds of dimension greater than two. From that point of view the role of the theory of surfaces is as a useful simple prototype. But there are a number of questions concerning surfaces in which interest is still very alive. One such is the theory of minimal surfaces (i.e. surfaces of minimal area spanning a given curve in space). Another is concerned with the "ergodic" aspects of geodesics: on many closed surfaces, but not on all, almost all geodesics eventually pass arbitrarily close to every point of the surface. The most important area of active research, however, is concerned with the study of the totality of possible metrics which a given surface can have: it turns out that this reduces to studying the surfaces from a holomorphic point of view, i.e. considering them as Riemann surfaces.

There are many books about the classical differential geometry of surfaces, all covering much the same ground. An excellent account, which is very thorough and goes far beyond the material here, can be found in

M.P. do Carmo, Differential Geometry of Curves and Surfaces (Prentice Hall)

It contains in particular a very good supply of suitable exercises.

I am most grateful to Glenys Luke and Wilson Sutherland for helpful comments on the manuscript of these notes.

§1 The definition of a surface, and some examples

Definition (1.1) A surface is a Hausdorff topological space which is locally homeomorphic to $\mathbb{R}^2$.

This definition requires some comments. First, to say that a space X is locally homeomorphic to $\mathbb{R}^2$ means that each point $x \in X$ is contained in an open set U which is homeomorphic to an open set V of $\mathbb{R}^2$.

We ask for the space to be Hausdorff to eliminate perverse examples which do not resemble our intuition of a surface. A space whose topology is defined by a metric is automatically Hausdorff. In all the examples of interest to us the topology can be defined by a metric. But we prefer to define a surface as a topological space rather than as a metric space because for many purposes the metric is irrelevant, and often there is no natural choice. (A good illustration is provided by Example (1.2) below.)

According to our definition, the following are surfaces (cf. Ex. 1.1):

(i) $\mathbb{R}^2$ itself

(ii) any open set of $\mathbb{R}^2$

(iii) the sphere $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$, a subspace of $\mathbb{R}^3$,

(iv) the surface of a cube in $\mathbb{R}^3$, and

(v) the cone $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = z^2 \text{ and } z > 0\}$.

The following, on the other hand, are not surfaces.

(i) the closed disc $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$, and

(ii) the double cone $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = z^2\}$;

for no neighbourhood of a boundary point of the disc, or of the vertex of the double cone, is homeomorphic to an open set of $\mathbb{R}^2$. (These facts are not so easy to prove. Cf. Ex.1.2)

So far we have mentioned only subsets of ordinary Euclidean space $\mathbb{R}^3$. But many examples of surfaces - probably the most important ones in the end - arise more abstractly.

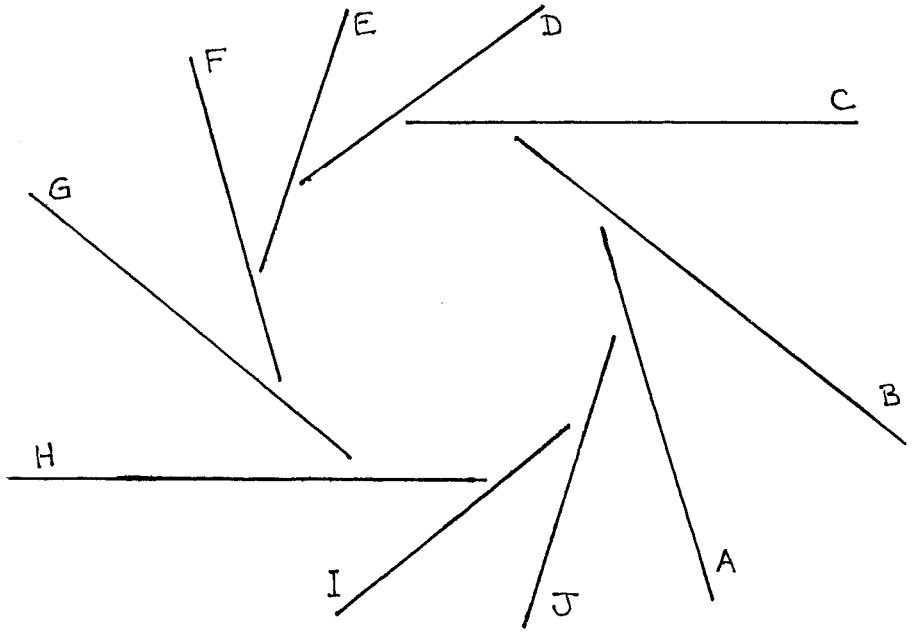
Example (1.2)

Let X be the set of all straight lines in the plane $\mathbb{R}^2$. Intuitively it is clear that X is a topological space : we have no doubts deciding whether a moving line is moving continuously or not, and we feel sure that the function $X \to \mathbb{R}$ which associates to a line its distance from the origin is continuous, whereas the function $X \to \mathbb{R}$ which associates to a line its slope (regarded as an angle in the half-open interval $(-\frac{\pi}{2}, \frac{\pi}{2}]$) is discontinuous.

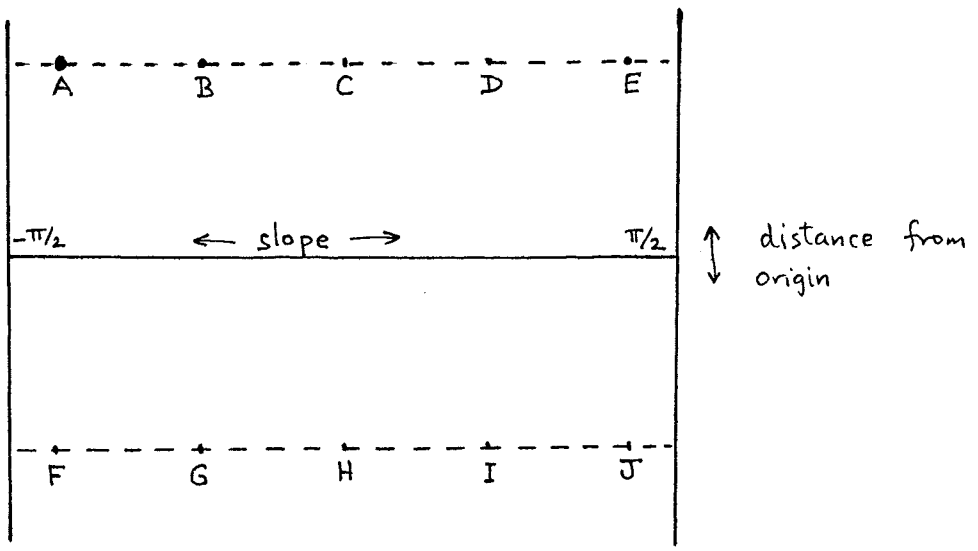
Let us try to make a picture of the space of lines. If we were content to leave out all the vertical lines then the task would be easy. Let X_o be the set of non-vertical lines. An element of X_o is specified by its slope $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$ and its signed distance from the origin 0 - taken positive if the line passes above 0, and negative if

it passes below 0. Thus it is very natural to identify X_o with the open subset $(-\frac{\pi}{2}, \frac{\pi}{2}) \times \mathbb{R}$ of $\mathbb{R}^2$. We now want to add the vertical lines, with slope $\pm\frac{\pi}{2}$, to the picture. We can add them as the left-hand edge $\{-\frac{\pi}{2}\} \times \mathbb{R}$ or as the right-hand edge $\{\frac{\pi}{2}\} \times \mathbb{R}$, or (preferably) as <u>both</u>, like the International Date Line on a conventional map of the world. But which point of $\{-\frac{\pi}{2}\} \times \mathbb{R}$ or $\{\frac{\pi}{2}\} \times \mathbb{R}$ should correspond to which vertical line? We can decide as follows.

Consider the sequence of lines A, B, C, ... J,



all at unit distance from the origin, and all belonging to X_o. They are represented in $(-\frac{\pi}{2}, \frac{\pi}{2}) \times \mathbb{R}$ by :



Thus the vertical line "between" E and F should be represented by either $(\frac{\pi}{2}, 1)$ or by $(-\frac{\pi}{2}, -1)$ on the picture, and the vertical line between A and J should be represented by $(\frac{\pi}{2}, -1)$ or $(-\frac{\pi}{2}, 1)$. For any $y \in \mathbb{R}$ the points $(\frac{\pi}{2}, y)$ and $(-\frac{\pi}{2}, -y)$ represent the same vertical line. To get a proper picture of X we must take $[-\frac{\pi}{2}, \frac{\pi}{2}] \times \mathbb{R}$ and attach the boundary lines to each other so that $(-\frac{\pi}{2}, -y)$ is identified with $(\frac{\pi}{2}, y)$ for all $y \in \mathbb{R}$. This gives us not a cylinder but a Möbius band.

So far our discussion has been heuristic. But we can now see how to define a topology on the set X so that it is indeed a surface. The preceding discussion showed us how to define a bijection a $\phi_o : X_o \to (-\frac{\pi}{2}, \frac{\pi}{2}) \times \mathbb{R}$. Let X_1 be the set of lines which are not horizontal. We can define a bijection $\phi_1 : X_1 \to (0,\pi) \times \mathbb{R}$ in the same way that we defined ϕ_o. We now make the definition that a subset U of X is open if $\phi_o(U \cap X_o)$ is an open subset of $(-\frac{\pi}{2}, \frac{\pi}{2}) \times \mathbb{R}$ and $\phi_1(U \cap X_1)$ an open subset of $(0,\pi) \times \mathbb{R}$. We leave it to the reader (Ex. 1.3) to check that this does define a topology which makes X a surface. This topology can be defined by a metric, but not in a very natural way. (Ex. 1.4)

Surfaces in $\mathbb{R}^3$

The most obvious surfaces are those defined by a single equation in $\mathbb{R}^3$, i.e. those of the form

$$X = \{(x, y, z) \in \mathbb{R}^3 : f(x, y, z) = 0\},$$

where $f : \mathbb{R}^3 \to \mathbb{R}$ is a continuously differentiable map. Not every f gives rise to a surface. We shall see in §5 that a sufficient condition for it to do so is that f and grad f do not vanish simultaneously.

Example

The three kinds of central quadric in $\mathbb{R}^3$ are

ellipsoids, with equation $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1 = 0,$

hyperboloids of one sheet, $\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} - 1 = 0,$

hyperboloids of two sheets, $\frac{x^2}{a^2} - \frac{y^2}{b^2} - \frac{z^2}{c^2} - 1 = 0.$

Where f and grad f do vanish simultaneously we expect X to behave badly. Thus if $f = x^2 + y^2 + z^2$ then f and grad f both vanish at the origin, and X consists of the origin alone, and is not a surface. If $f = x^2 + y^2 - z^2$ then f and grad f again vanish at the origin, and the surface is a double cone.

Two classes of surfaces that will be useful for illustrating results in differential geometry are :

(i) Surfaces of revolution, obtained by taking a curve γ in the XZ plane and rotating it about the Z axis. (One must choose a curve γ which either does not meet the Z-axis or else is symmetric about it.) An important case is the torus, got by rotating the circle

$$(x-b)^2 + z^2 = a^2$$

with centre (b,0) and radius a, where $b > a$.

(ii) Ruled surfaces, which are swept out by a straight line moving in $\mathbb{R}^3$. The hyperboloid of one sheet is of this kind; in fact it is a ruled surface in two different ways, and is the only surface with that property.(*) (Cf. Exercises 1.5, 1.6, 1.7.)

An important subclass of the ruled surfaces are the developables, which are swept out by the tangent line to a curve in space. We shall see in §7 that these are the most general surfaces which can be obtained from a piece of a plane by bending it without stretching it.

Note One usually has to exclude some "bad" set from the locus of a moving line to obtain a surface. Thus the double cone $x^2 + y^2 = z^2$ is swept out by a line, but is a surface only if one omits the origin. The developable surface swept out by the tangents to a curve γ has a sharp edge (called a "cuspidal edge") along γ itself, and one must exclude that to have a surface.

Complex algebraic curves

A complex algebraic curve in $\mathbb{C}^2$ is a set of the form

$$X = \{(x, y) \in \mathbb{C}^2 : f(x, y) = 0\},$$

where f is a polynomial in two variables with complex coefficients. Because $f = 0$ amounts to two real equations in four real variables we expect X to be a surface. That is true providing f and grad f do not vanish simultaneously. (See the Appendix.)

(*) Apart from the hyperbolic paraboloid (e.g. $z = xy$), which is a limiting case of the hyperboloid.

Example (1.3)

The equation $x^2 + y^2 = 1$ defines a surface X in $\mathbb{C}^2$. If we write the complex vector $\binom{x}{y}$ as u + iv, where u and v belong to $\mathbb{R}^2$, then the equation becomes the pair of equations

$$\|u\|^2 - \|v\|^2 = 1$$
$$\langle u, v \rangle = 0$$

(Thus $u \neq 0$.) Each solution of these equations can be written in the form

$$u = \xi \cosh t$$
$$v = \xi^{\perp} \sinh t, \qquad (1.4)$$

where ξ is the unit vector $u/\|u\|$ in $\mathbb{R}^2$, $\xi^{\perp}$ is the vector obtained by rotating ξ through $\pi/2$, and $t \in \mathbb{R}$. Conversely, for each ξ on the unit circle in $\mathbb{R}^2$ and each $t \in \mathbb{R}$ the formulae (1.4) define a point of X. Thus X is topologically a cylinder, the cartesian product of the circle $\|\xi\| = 1$ and the line $\mathbb{R}$.

Another way to see the topological type is to parametrize X by

$$x = \frac{1+w^2}{2w}, \qquad y = \frac{1-w^2}{2iw},$$

where $w = (x+iy)^{-1} \in \mathbb{C} - \{0\}$. This shows that X is homeomorphic to $\mathbb{C} - \{0\}$, which is topologically a cylinder because each $w \in \mathbb{C} - \{0\}$ can be written uniquely $e^t u$ with $t \in \mathbb{R}$, $u \in \mathbb{C}$, and $|u| = 1$.

The following rather vague remark is designed to help introduce some future ideas. In complex variable theory it is often convenient to adjoin to the complex plane one extra point called ∞ : the set $\mathbb{C} \cup \{\infty\}$ is called the Riemann sphere. In the same way it is natural to adjoin two "points at infinity" to the surface X. We can call them $P_1 = (\infty, i\infty)$ and $P_2 = (\infty, -i\infty)$. They correspond to the parameter values $w = \infty$ and $w = 0$. Thus $X \cup \{P_1, P_2\}$ is in 1-1 correspondence with the Riemann sphere. Geometrically P_1 and P_2 are the "ends" of the two asymptotes $x \pm iy = 0$ of the complex curve $x^2 + y^2 = 1$: these asymptotes are exactly analogous to the usual real asymptotes $x \pm y = 0$ of the hyperbola $x^2 - y^2 = 1$ in $\mathbb{R}^2$. If $x^2 + y^2 = 1$ and $|x|$ is large then $y = i\sqrt{x^2-1}$ is either very close to ix or very close to $-ix$.

Exercises

1.1 Prove that the examples (iii), (iv), (v) of surfaces on page 5 really are surfaces.
[In case (iv) it is helpful to begin by proving the following lemma. If a space X is the union of a finite number of closed subsets X_i, then a map $f : X \to Y$ is continuous if the restriction of f to X_i is continuous for each i.]

1.2 Prove that the double cone
$X = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = z^2\}$ is not a surface.
[Observe that X becomes disconnected if the origin is deleted.]

1.3 Prove that the definition of open sets given on page 8 does define a topology on the set of lines in $\mathbb{R}^2$, and makes it a surface.

1.4 (a) Find a metric on the set X of lines in $\mathbb{R}^2$ which defines its topology.

(b) Prove that there is no metric d on X which defines its topology and has the property that $d(\ell_1, \ell_2) = d(T(\ell_1), T(\ell_2))$ for all $\ell_1, \ell_2 \in X$ and every rigid motion $T : \mathbb{R}^2 \to \mathbb{R}^2$. [If such a metric existed then the distance between two lines which intersect at an angle θ would depend only on θ. Call the distance $f(\theta)$. Prove $f(\theta) \to 0$ as $\theta \to 0$. Then contradict the triangle inequality by considering a pair of parallel lines and a third line which intersects them both at a small angle.]

1.5. Show that for every non-zero $\lambda \in \mathbb{R}$ the straight line with equations $x-z = \lambda(1-y)$, $x+z = \lambda^{-1}(1+y)$ lies on the hyperboloid $x^2 + y^2 - z^2 = 1$. Deduce that every hyperboloid of one sheet is a ruled surface. Find another family of lines on $x^2 + y^2 - z^2 = 1$, and show that lines of the same family do not intersect, but that each line of the first family meets each line of the second.

1.6 Prove that if a straight line is rotated rigidly about a line not in the same plane, then it sweeps out a hyperboloid of one sheet.

1.7. Let ℓ_1, ℓ_2, ℓ_3 be three lines in $\mathbb{R}^3$, no two coplanar. Prove that through each point P of ℓ_3 there is a unique line ℓ_P which meets both ℓ_1 and ℓ_2. Prove that as P varies on ℓ_3 the line ℓ_P sweeps out a hyperboloid of one sheet. Deduce that this surface is the only surface which is ruled in two different ways. [It requires ingenuity to do this question from first principles. We shall encounter better methods in the second half of this course. The equation of any line in $\mathbb{R}^3$ can be written in the form $a \times r = b$. Prove that $a \times r = b$ intersects $c \times r = d$ if and only if $\langle a,d\rangle + \langle b,c\rangle = 0$. Deduce that there is a line $c \times r = d$ through r which meets each of $a_i \times r = b$, for $i = 1, 2, 3$ if and only if the scalar triple product

$$[a_1 \times r - b_1,\quad a_2 \times r - b_2,\quad a_3 \times r - b_3]$$

vanishes. Show that this is the equation of a quadric surface, necessarily a hyperboloid of one sheet.]

1.8 A ruled surface X is swept out by the line through the point $\gamma(t)$ in the direction of the unit vector $a(t)$, where γ and a are continuously differentiable maps $(\alpha,\beta) \to \mathbb{R}^3$. Assuming that $\dot{a} = da/dt$ does not vanish for $t \in (\alpha,\beta)$, prove that X is developable if and only if the scalar triple product $\langle \dot{\gamma}, a \times \dot{a}\rangle$ is zero.

[The problem is to find a curve $\rho(t) = \gamma(t) + f(t)a(t)$ on X such that the tangent $\dot{\rho}$ is parallel to a. If $\dot{\rho} = ga$ one must find f and g from $ga = \dot{\gamma} + f\dot{a} + \dot{f}a$.]

§2 Abstract surfaces

The surfaces which are important in mathematics mostly do not appear as objects in space. In this section, we shall give some examples of ways in which surfaces arise abstractly. Nothing in this section is essential to the course, and it will not be referred to again, so readers who find it confusing can simply omit it.

A. The torus and the Klein bottle as quotient spaces.

A typical way in which a torus arises is as the set of positions of the hands of a clock. A position of the hands is a pair (x,y), where each of x and y is a real number modulo 12. Thus the set X of positions is obtained from the plane $\mathbb{R}^2$ by introducing the equivalence relation $\sim$ such that $(x_1,y_1) \sim (x_2,y_2)$ if and only if

$$\begin{aligned} x_1 &= x_2 + 12.n \\ y_1 &= y_2 + 12.m \end{aligned} \qquad \text{for some integers } n,m. \tag{2.1}$$

We want to think of the set of positions as a topological space. We know the topology of $\mathbb{R}^2$, and there is an obvious map $\mathbb{R}^2 \to X$ which assigns to each $(x,y) \in \mathbb{R}^2$ its equivalence class in X. We define an open set of X as one whose inverse-image in $\mathbb{R}^2$ is open. A little reflection convinces one that this agrees with our intuitive idea of what an open set in X ought to be. We can now prove that X is homeomorphic to a standard torus in $\mathbb{R}^3$. (Ex. 2.1)

What we have just described is a method which gives us a topology on any set which is the set of equivalence classes of an equivalence relation on another topological space. The topology constructed is called the quotient topology.

Now let us consider a more subtle equivalence relation on $\mathbb{R}^2$. We define $(x_1,y_1) \sim (x_2,y_2)$ if and only if

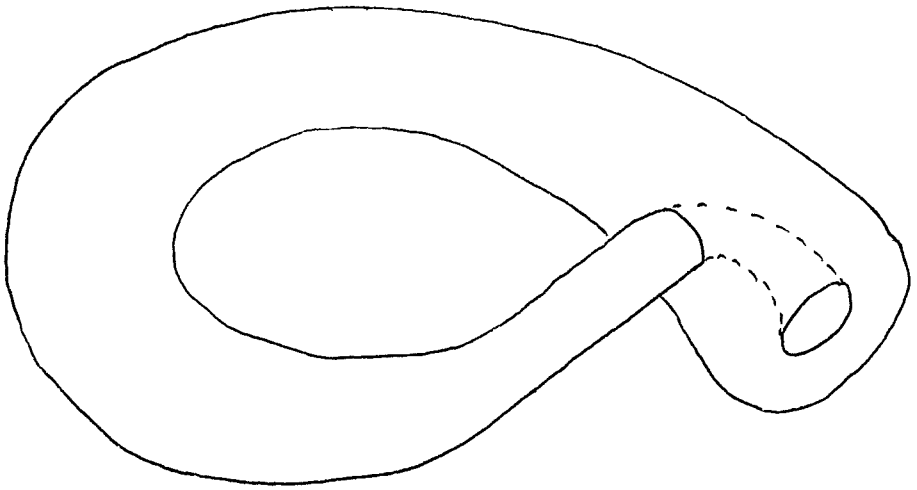
$$\begin{aligned} x_1 &= x_2 + n \\ y_1 &= (-1)^n y_2 + m \end{aligned} \qquad \text{for some integers } n,\ m. \tag{2.2}$$

Then every point of $\mathbb{R}^2$ is equivalent to a point in the square $[0,1] \times [0,1]$, and no two points of the square are equivalent except that

$$(x,\ 0) \sim (x,\ 1) \quad \text{for } 0 \leqslant x \leqslant 1, \text{ and}$$
$$(0,\ y) \sim (1,\ 1-y) \quad \text{for } 0 \leqslant y \leqslant 1.$$

The space of equivalence classes $X = \mathbb{R}^2/\sim$ is easily checked to be a surface. (Ex. 2.2) It is called the Klein bottle. (*) Unlike the torus it cannot be realized as a surface in $\mathbb{R}^3$, although we can find a map $f : X \to \mathbb{R}^3$ which is locally a homeomorphism on to $f(X)$ but which is not quite 1-1 : the space $f(X)$ is the usual picture of the Klein bottle, complete with "self-intersection".

(*) It is not hard to envisage a contrivance whose set of positions is this surface. Suppose a device consists of a plane Π in $\mathbb{R}^3$ free to rotate about the Z-axis, together with a line ℓ through the origin constrained to lie in the plane Π. Then the positions of (Π,ℓ) form a Klein bottle.

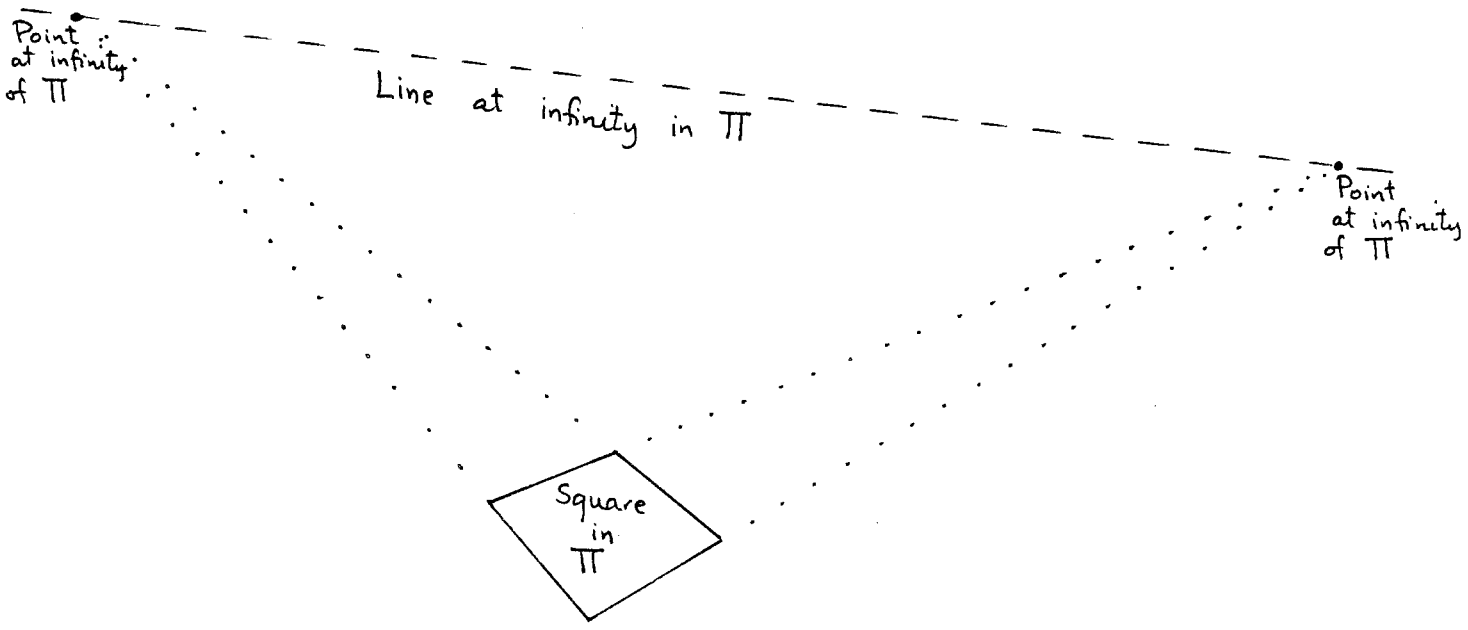


B. The projective plane

A very important abstract surface is the projective plane, which will be studied in detail in the second half of the course. The idea of its construction is as follows. Imagine one is at the origin in $\mathbb{R}^3$, looking at the plane Π whose equation is $z = 1$. There is almost a 1-1 correspondence between the points of the plane Π and the rays to one's eye, i.e. between Π and the set of lines through the origin in $\mathbb{R}^3$. But the correspondence is not perfect, because lines parallel to Π, i.e. those in the XY-plane, do not meet Π. As a ray becomes more nearly parallel to Π its point of intersection moves away "towards infinity". It is therefore tempting to adjoin to Π a set Π_∞ of "ideal" points "at infinity", one for each ray in the XY plane. The combined set $\hat{\Pi} = \Pi \cup \Pi_\infty$ is called the projective plane.

The idea of introducing "points at infinity" originates in perspective drawing, when one makes a picture of a plane Π in $\mathbb{R}^3$ on another plane Π' which is not parallel to it. Then the "points at infinity" of Π are depicted by ordinary points of Π' - so-called "vanishing points" of families of

parallel lines in Π.



Picture of a plane Π on the plane Π' of the paper

If we add the "points at infinity" to both Π and Π' then there is an exact 1-1 correspondence between them.

The formal definition of the projective plane is psychologically unilluminating.

Definition (2.3) The projective plane is the set of lines through the origin in $\mathbb{R}^3$.

A line through the origin is determined by giving a single non-zero vector v on it; and v and v' determine the same line if and only if $v' = \lambda v$ for some non-zero scalar λ. Thus we can reformulate the definition as follows.

Definition (2.4) The projective plane is the set of equivalence classes of $\mathbb{R}^3 - \{0\}$ for the equivalence relation defined by

$$v \sim v' \iff v' = \lambda v \text{ for some } \lambda \neq 0 \text{ in } \mathbb{R}.$$

An advantage of the second version of the definition is that it makes clear that the projective plane is a topological space.

According to the second definition, a point of the projective plane $\hat{\Pi}$ is described by homogeneous coordinates (x, y, z), not all zero, subject to the convention that (x, y, z) and $(\lambda x, \lambda y, \lambda z)$ describe the same point if $\lambda \neq 0$. To prove that $\hat{\Pi}$ really is a surface we shall show that it is the union of three open sets U_1, U_2, U_3 each of which is homeomorphic to $\mathbb{R}^2$. We define U_1 as the set of all points whose homogeneous coordinates (x, y, z) satisfy $x \neq 0$. Similarly U_2 and U_3 consist of points such that $y \neq 0$ and $z \neq 0$ respectively. We have already agreed to identify U_3 with the plane $z = 1$ by

$$(x, y, z) \longleftrightarrow \left(\frac{x}{z}, \frac{y}{z}, 1\right),$$

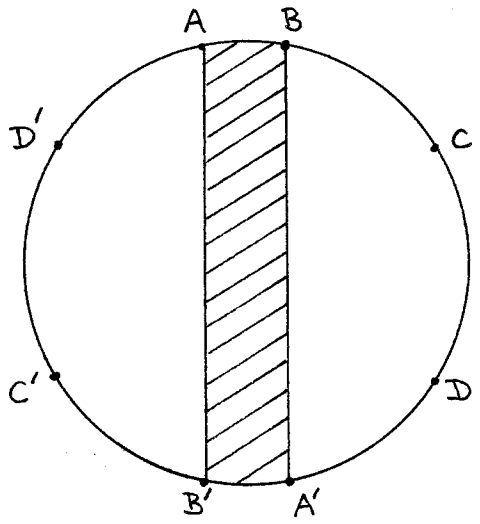
and one can check that this is a homeomorphism. (Ex. 2.3) Similarly U_1 and U_2 can be identified with the planes $x = 1$ and $y = 1$.

To visualize the projective plane it is best to start from another topological description of it, whose validity we leave as an exercise. (Ex. 2.4)

Theorem (2.5) The projective plane is the quotient space of the disc $\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ by the equivalence relation which identifies opposite points of the boundary:

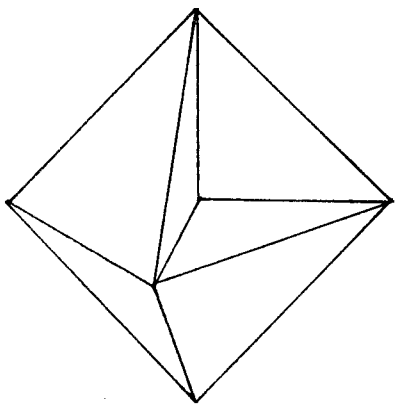
$$(x,y) \sim (x',y') \iff x = x' \text{ and } y = y'$$
$$\text{or} \quad x = -x', \; y = -y', \text{ and } x^2 + y^2 = 1.$$

Like the Klein bottle, the projective plane cannot be realized as a surface in $\mathbb{R}^3$ without self-intersections. Notice that the shaded part in the diagram ia a Möbius band. The unshaded part fits together to form a disc; thus $\hat{\Pi}$ is the union of a disc and a Möbius band.

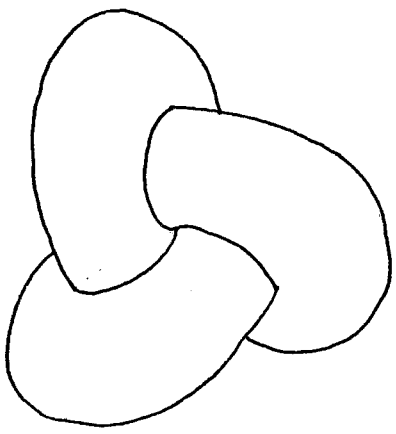


$A \sim A'$
$B \sim B'$
$C \sim C'$
$D \sim D'$

One possible self-intersecting representation of $\hat{\Pi}$ in $\mathbb{R}^3$ is as a heptahedron, which is made from three squares with unit sides intersecting perpendicularly along their diagonals by adding four equilateral triangles also with unit sides. The most beautiful representation of the projective plane, however, is as Boy's surface.



Heptahedron



Boy's surface

C. Riemann surfaces

Our last class of examples of abstract surfaces is the hardest to motivate, but is ultimately the most important.

In complex analysis one constantly encounters "many-valued functions" such as $\log z$ and $\sqrt{1-z^2}$. The most elementary way of dealing with them is to restrict oneself to an open set V of the complex plane in which one can define a single-valued holomorphic function which at each point takes "one of the values" of the ill-defined function one is interested in. In the case of $\log z$, for example, one can take V to be $\mathbb{C}$ with the negative real axis removed, and in V one can define

$$\log (re^{i\theta}) = \log r + e^{i\theta},$$

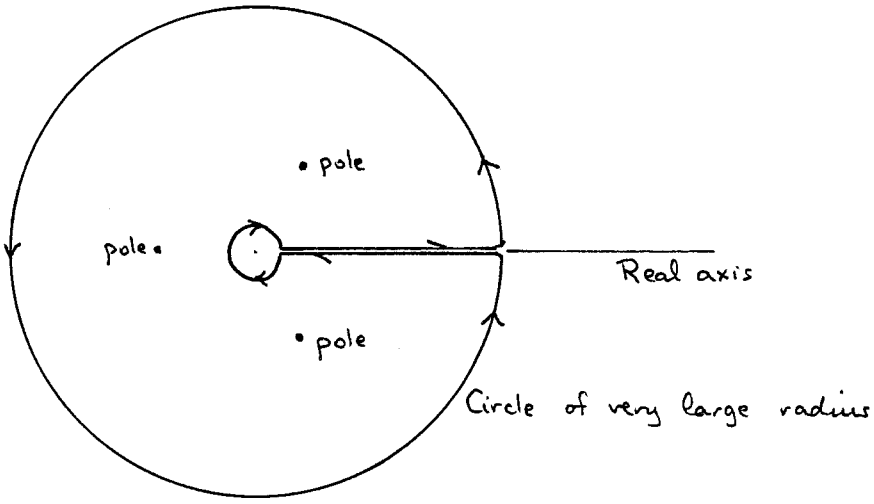
with $-\pi < \theta < \pi$. This approach is adequate for many purposes, but it has some disadvantages. One is that the choice of V is rather arbitrary - we could have made the cut along the negative imaginary axis, for example - and $\log z$ behaves just as well or as badly at any point of $\mathbb{C} - \{0\}$ as at any other.

A simple example which shows the beginnings of the idea of a Riemann surface is the following. The quickest way to evaluate

$$\int_0^\infty \frac{dx}{1+x^3} = \frac{2\pi\sqrt{3}}{9}$$

is to integrate the function $\log z \,/\, (1+z^3)$ around the

"key-hole" contour γ :



We take the definition of log z, everywhere except on the positive real axis, to be $\log|z| + i \arg z$, with $0 < \arg z < 2\pi$. But the contour involves integrating twice along the real axis, once using the real values $\log |z|$, and once using $\log|z| + 2\pi i$. We can avoid the problem by displacing the contour slightly away from the cut and taking a limit, but that introduces an unnecessary complication. A better, though more sophisticated, approach is to introduce the Riemann surface on which the function $\log z / (1+z^3)$ is defined. We take a stack $\{X_k\}$ of copies of the complex plane, each cut along the positive real axis. Then we attack the lower lip of the cut in X_k to the upper lip of the cut in X_{k+1} for each k, so as to obtain a surface X which is like an infinite spiral staircase squashed flat. The function $\log z / (1+z^3)$ is a genuine function on X : on X_k we have

$$2\pi k \leqslant \mathrm{Im}(\log z) \leqslant 2\pi(k+1).$$

The contour γ lies naturally on X : the two transits of the real axis lie on distinct sheets of X.

Now let us consider the Riemann surface for the function

$\sqrt{1-z^2}$. Generically the function is two-valued, so we take two copies X^+ and X^- of $\mathbb{C}$, each cut from -1 to +1. Then we attach the upper lip of X^+ to the lower lip of X^- and the lower lip of X^+ to the upper lip of X^-. We define $\sqrt{1-z^2}$ on $X = X^+ \cup X^-$ so that $\mathrm{Im}\sqrt{1-z^2} \geqslant 0$ for $z \in X^+$ and $\mathrm{Im}\sqrt{1-z^2} \leqslant 0$ for $z \in X^-$. (Notice that $\sqrt{1-z^2}$ is real only if $z \in [-1,1]$). In this case, unlike that of log z, the function $\sqrt{1-z^2}$ takes the well-defined value 0 at each of the branch-points $z = \pm 1$, so one can add these two points to X. The surface X is at first hard to visualize, but it is homeomorphic to $\mathbb{C} - \{0\}$ by the map which takes $\zeta \in \mathbb{C} - \{0\}$ to $\frac{1}{2}(\zeta+\zeta^{-1}) \in X^+$ if $|\zeta| \geqslant 1$, and to $\frac{1}{2}(\zeta+\zeta^{-1}) \in X^-$ if $|\zeta| \leqslant 1$. (See Ex. 2.5.)

The abstract surface which we call the Riemann surface of the multivalued function f is the same thing as the graph of f, i.e. the set of all pairs $(z,w) \in \mathbb{C}^2$ such that w is one of the values of f(z). Thus the Riemann surface of $\sqrt{1-z^2}$ is the same thing as the complex algebraic curve $x^2 + y^2 = 1$ which we described in §1. It is a matter of taste whether one prefers to regard it as a subset of $\mathbb{C}^2$ or as a collapsed parachute lying on the complex plane, but it can be useful to move between the two pictures in one's mind. We shall meet an illustration of this at the end of §4.

Exercises

In doing the following exercises it is useful to remember the following obvious principle: to define a continuous map $f : X \to Y$, where the topological space X is defined as the quotient space of a space $\tilde{X}$ by an equivalence

relation $\sim$, it is enough to define a continuous map $\tilde{f} : \tilde{X} \to Y$ which is compatible with the equivalence relation, i.e. is such that $\tilde{f}(\tilde{x}) = \tilde{f}(\tilde{x}')$ whenever $\tilde{x} \sim \tilde{x}'$.

2.1 If X is the quotient space of $\mathbb{R}^2$ by the equivalence relation (2.1), prove that X is homeomorphic to a standard torus Y in $\mathbb{R}^3$, and also homeomorphic to $\{(z_1,z_2) \in \mathbb{C}^2 : |z_1| = |z_2| = 1\}$
[Define a continuous map $\mathbb{R}^2 \to Y$ which induces a continuous map $f : X \to Y$ by the principle above. Prove that f is a bijection. Prove that X is compact because it is the image of a compact subspace of $\mathbb{R}^2$. Finally, use the theorem that a continuous bijection from a compact space to a Hausdorff space is a homeomorphism.]

2.2 If X is the quotient space of $\mathbb{R}^2$ by the equivalence relation (2.2), prove that X is a surface.
[If $x \in X$ is the equivalence class of $v \in \mathbb{R}^2$, and $V = \{v' \in \mathbb{R}^2 : \|v'-v\| < \frac{1}{2}\}$, prove that the obvious map $V \to X$ is a homeomorphism between V and a neighbourhood of x.]

2.3 If U_3 is the subset of the projective plane $\hat{\Pi}$ described on page 19, prove that U_3 is an open subset of $\hat{\Pi}$, and that the map $(x, y, z) \longmapsto (x/z, y/z)$ defines a homeomorphism $U_3 \to \mathbb{R}^2$.
[Define the inverse map as a composite : $\mathbb{R}^2 \to \mathbb{R}^3 - \{0\} \to \hat{\Pi}$.]

2.4 Prove Theorem (2.5).
[Begin with the continuous map $\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant 1\} \to \hat{\Pi}$ defined by $(x,y) \longmapsto (x, y, 1 - x^2 - y^2)$, and proceed as in Ex. 2.1.]

2.5 Justify the assertion on page 23 that the Riemann surface X is homeomorphic to $\mathbb{C} - \{0\}$ by showing that $\zeta \longmapsto \frac{1}{2}(\zeta + \zeta^{-1})$ is a bijection from $\{\zeta \in \mathbb{C} : |\zeta| \geqslant 1\}$ to X^+ which takes the semicircle $C^+ = \{\zeta \in \mathbb{C} : |\zeta| = 1 \text{ and } \operatorname{Im} \zeta \geqslant 0\}$ to the upper lip of the cut and the semicircle $C^- = \overline{C}^+$ to the lower lip of the cut; while the same map is also a bijection from $\{\zeta \in \mathbb{C} : 0 < |\zeta| \leqslant 1\}$ to X^- which takes C^+ to the lower and C^- to the upper lip of the cut.

§3 Charts, atlases, orientability

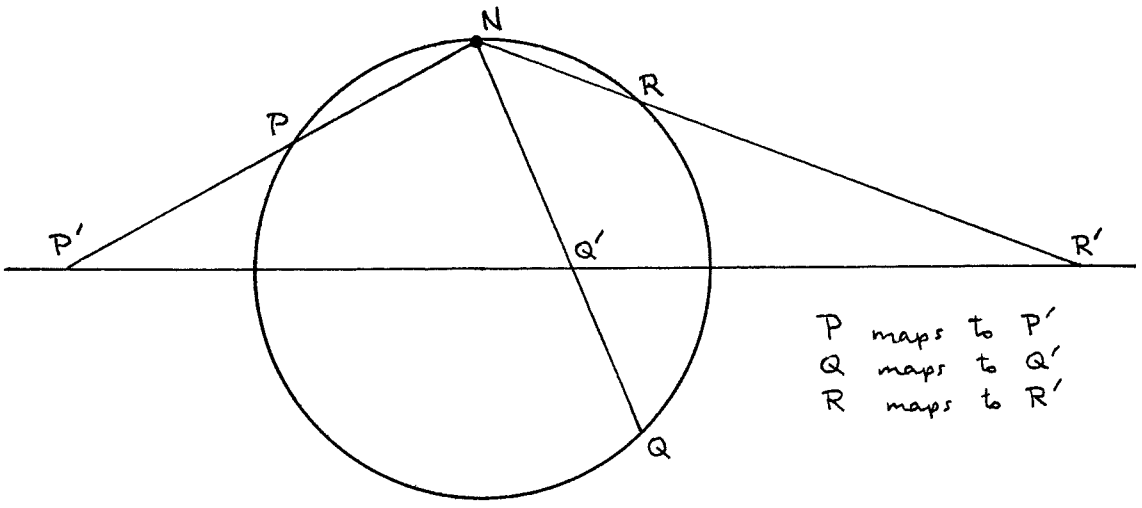
If X is a surface, a homeomorphism $\phi : U \to V$ from an open set U of X to an open set V of the plane $\mathbb{R}^2$ is called a chart (or coordinate system) for X. A collection of charts $\{\phi_\alpha : U_\alpha \to V_\alpha\}$ such that the sets U_α cover X is called an atlas for X. By definition, every surface possesses an atlas.

If X is a surface in $\mathbb{R}^3$, the most obvious way to produce charts for it is by projection on to the coordinate planes of $\mathbb{R}^3$. Thus if X is the unit sphere in $\mathbb{R}^3$, and U is the open hemisphere defined by $z > 0$, the map $(x, y, z) \mapsto (x, y)$ from U to the open unit disc in $\mathbb{R}^2$ is a chart, and X has an atlas consisting of six such charts, each covering an open hemisphere.

Charts defined by projection are seldom the ones which are useful in practice. Thus for the sphere the best known chart is the one given by longitude and latitude. It is defined on the open set U got by removing the International Date Line from the sphere, and it is a homeomorphism between U and the open set $(-\pi,\pi) \times (-\frac{\pi}{2}, \frac{\pi}{2})$ of $\mathbb{R}^2$. Even this chart, however, is not usually found in a geographical atlas. Granted that one cannot make a chart for a region of the earth which is exactly to scale (*) geographers want charts which have some useful but weaker property. Three such properties are (i) conformality, i.e. preserving angles, (ii) preserving areas, and (iii) representing great circles (which are the shortest routes on the sphere) by straight lines. Let us mention examples of each kind.

(*) Although this is well known, we shall not be able to prove it before §10.

(i) Stereographic projection is the chart $\phi : U \to V$ given by projection from the north pole on to the equatorial plane.



Thus, for the unit sphere $x^2 + y^2 + z^2 = 1$, the domain U is $X - \{(0,0,1)\}$, and V is $\mathbb{R}^2$, and

$$\phi(x, y, z) = \left(\frac{x}{1-z}, \frac{y}{1-z}\right).$$

This is a conformal chart (Ex. 3.3), as is well-known in complex variable theory.

Mercator's projection is defined on the complement of the Date Line. It takes the point with longitude ϕ and latitude θ to

$$(\phi, \log \tan(\frac{\theta}{2} + \frac{\pi}{4})) \quad \in \quad (-\pi, \pi) \times \mathbb{R}.$$

This is conformal (Ex. 3.4), and was especially useful for navigation, and for depicting the British Empire in the days of its glory.

(ii) The most obvious area-preserving chart (Ex. 3.5) takes (ϕ,θ) to $(\phi, \sin\theta)$. (Here (ϕ,θ) are longitude and latitude again).)

A more popular one is Mollweide's projection, which takes (ϕ,θ) to $(\phi\cos\psi(\theta), \tfrac{1}{2}\pi \sin\psi(\theta))$, where $\psi : [-\tfrac{\pi}{2},\tfrac{\pi}{2}] \to [-\tfrac{\pi}{2},\tfrac{\pi}{2}]$ is the bijection defined by

$$\psi(\theta) + \tfrac{1}{2} \sin 2\psi(\theta) = \tfrac{1}{2}\pi \sin\theta.$$

(iii) Essentially the only chart which takes great circles to straight lines is projection from the centre of the earth on to a tangent plane. Thus we can map the open northern hemisphere to $\mathbb{R}^2$ by

$$(\phi,\theta) \longmapsto (\cot\theta\cos\phi\ ,\ \cot\theta\sin\phi).$$

Orientability

It is a basic fact of nature that homeomorphisms $f : V \to V'$ from one connected open set of $\mathbb{R}^2$ to another come in two kinds: orientation-preserving and orientation-reversing. The first kind take clockwise simple closed curves to clockwise ones, the second kind take clockwise curves to anticlockwise ones. For the moment we shall simply accept the existence of this dichotomy. (Cf. §5.)

We can now divide surfaces into two classes, orientable and non-orientable. To do this, first observe that if $\phi : U \to V$ and $\phi' : U' \to V'$ are two charts for the same surface then we have a homeomorphism.

$$\phi'\circ\phi^{-1} : \phi(U\cap U') \to \phi'(U\cap U')$$

between the two open sets of $\mathbb{R}^2$ which are the maps of $U\cap U'$. This homeomorphism is called the transition map from the first chart to the second.

Definition (3.1) A surface is orientable if it possesses an atlas for which all the transition maps are orientation-preserving. (*)

Of the surfaces we have encountered so far, all are orientable except the Möbius band, the Klein bottle, and the projective plane. To prove the negative statements it is enough to consider the Möbius band, for an open set of an orientable surface is obviously orientable, and the Klein bottle and the projective plane each contain Möbius bands. For the Möbius band see Ex. 3.7.

The question of orientability is very closely connected with whether the surface has one or two sides. But the latter question refers to a surface embedded in $\mathbb{R}^3$ in a definite way, whereas orientability is an intrinsic property of the surface as a topological space. We shall discuss one- and two-sidedness in §5.

The definition of a surface by means of an atlas

The Riemann sphere Σ is the set of complex numbers together with another element which is called ∞. Thus $\Sigma - \{\infty\} = \mathbb{C}$, and there is also a bijection $\phi : \Sigma - \{0\} \to \mathbb{C}$ given by

$$z \longmapsto z^{-1} \qquad \text{if } z \neq \infty,$$

$$\infty \longmapsto 0.$$

(*) A map between open sets of $\mathbb{R}^2$ is called orientation-preserving if it is orientation-preserving on each connected component.

We define the topology of Σ by saying that U is an open set of Σ if both $U - \{\infty\}$ and $\phi(U - \{0\})$ are open sets of $\mathbb{C}$. It is then easy to prove that Σ is homeomorphic to the unit sphere in $\mathbb{R}^3$ by stereographic projection.

It will be seen that we have here a general method for defining surfaces. (We have already used it for the Möbius band in §1.) If X is a set which is the union of a family of subsets $\{U_\alpha\}$, and for each α we are given a bijection $\phi_\alpha : U_\alpha \to V_\alpha$, where V_α is an open set in $\mathbb{R}^2$, then we can always define a toplogy on X by prescribing

$$U \text{ is open} \iff \phi_\alpha(U \cap U_\alpha) \text{ is open for all } \alpha.$$

Then X is a surface (see Ex. 3.9) providing

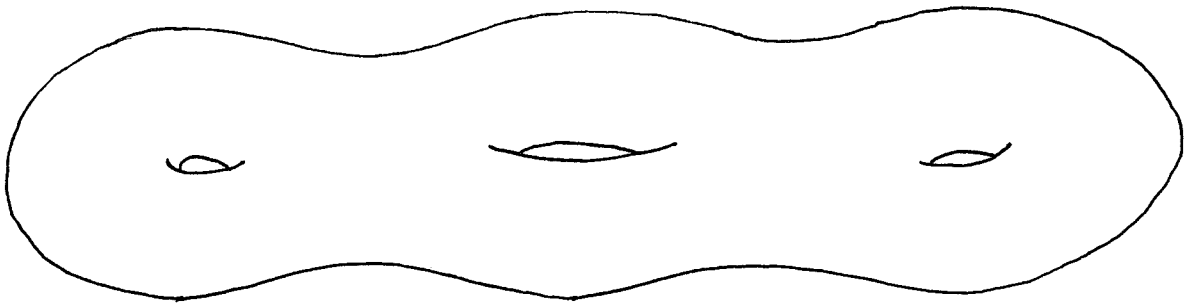
(a) each transition map $\phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta)$

is a homeomorphism, and

(b) $\{(x,x) : x \in U_\alpha \cap U_\beta\}$ is a closed subset of $U_\alpha \times U_\beta$.

The classification of surfaces

A surface which is a compact topological space is called a closed surface. It is a remarkable theorem that there are very few of them. We have already met the sphere and the torus, and it is easy to imagine a torus with g holes, for any integer $g \geqslant 0$:



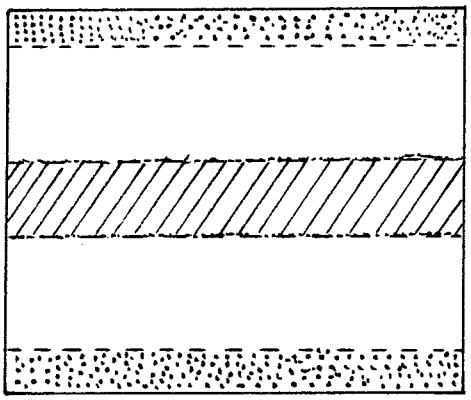
(We regard a sphere as a torus with 0 holes.)

Theorem (3.2) Any orientable closed surface is homeomorphic to a torus with g holes, for some $g \geqslant 0$.

Among non-orientable closed surfaces we have met the projective plane and the Klein bottle. A general method for constructing a non-orientable surface is the following. Observe that a Möbius band is a surface bounded by a single circle, just as a disc is. Then take a sphere, remove from it k disjoint discs, and replace each disc with a Möbius band. The resulting surface is called a sphere with k cross-caps.

Theorem (3.3) Any non-orientable closed surface is homeomorphic to a sphere with k cross-caps, for some $k \geqslant 1$.

The case k = 1 is the projective plane, as we have seen. The Klein bottle is the case k = 2. To see that, imagine the bottle made from a square by identifying edges:



The hatched part is a Möbius band; the dotted part is another Möbius band; and the remainder is a cylinder, i.e. a sphere with two discs removed.

We shall not prove Theorems (3.2) and (3.3) in the course.

It is much more complicated to classify non-closed surfaces. In particular it is not true that every surface is homeomorphic to an open set of a closed surface: consider, for instance, an infinite ladder made of tubular steel.

An atlas for the complex curve $x^n + y^n = 1$

Even the simplest functions of complex variables lead to surprisingly complicated surfaces. We shall take as an example the surface X in $\mathbb{C}^2$ defined by the equation $x^n + y^n = 1$, where n is a positive integer. In §4 we shall prove (using Theorem (3.2)) the striking result that this surface, together with its n "points at infinity" is a torus with $\frac{1}{2}(n-1)(n-2)$ holes. We shall now define an atlas for the surface: this is fairly complicated, and readers who find it confusing can ignore it. We use 2n charts to cover X itself, and n more to include the points at infinity.

For each x in the complex plane (except for the n roots of unity $e^{2\pi ik/n}$) there are n points (x,y) on X, for y can be any of the values of $\sqrt[n]{1-x^n}$. In order to define $\sqrt[n]{1-x^n}$ as a continuous function we have to cut the x-plane: we let V be the complex plane cut radially outwards from each of the n^{th} roots of unity to ∞, i.e.

$$V = \mathbb{C} - \{x : x^n \in \mathbb{R} \text{ and } x^n \geqslant 1\}.$$

If $x \in V$ then $1-x^n$ is not on the negative real axis, so we can define a holomorphic function $f_o : V \to \mathbb{C}$ by the conditions

$$f_o(x)^n = 1-x^n$$

and

$$-\pi/n < \arg f_o(x) < \pi/n.$$

We can also define the <u>other</u> branches of $\sqrt[n]{1-x^n}$, namely $f_k : V \to \mathbb{C}$, where $f_k(x) = e^{2\pi ik/n} f_o(x)$, for $o \leqslant k < n$.

Let U_k be the graph of the function f_k, i.e. the set

$$\{(x, f_k(x) \in \mathbb{C}^2 : x \in V\}.$$

This is a subset of X, and there is a homeomorphism $\phi_k : U_k \to V$ defined by $\phi_k(x,f_k(x)) = x$. Every point (x,y) of X such that $x \in V$ belongs to one of the sets U_k, and we therefore have n charts which cover <u>most</u> of X. We must now produce another n charts which cover the cracks. That is easy to do, for if $x^n + y^n = 1$ then x^n and y^n cannot <u>both</u> be real and $\geqslant 1$, i.e. at least one of x and y belongs to V. So we define

$$U_k' = \{(f_k(y), y) \in \mathbb{C}^2 : y \in V\},$$

and a homeomorphism $\phi_k' : U_k' \to V$ by $\phi_k'(f_k(y),y) = y$.

We now have an atlas for X.

Finally, we add the points at infinity. If $(x,y) \in X$ and $|x|$ is large then y is very near to $\omega_k x$, for some k, where $\omega_1,\ldots,\omega_n$ are the n^{th} roots of -1. (For

$$\begin{aligned} y &= (1-x^n)^{1/n} \\ &= (-1)^{1/n} x (1 - 1/x^n)^{1/n} \\ &= (-1)^{1/n} x (1 - 1/nx^n + \ldots) \end{aligned}$$

Thus X has n asymptotes, given by $y = \omega_k x$, and it is reasonable to adjoin n points $P_1,\ldots, P_n$ "at infinity", one at the end of each asymptote. (We think of P_k as "$(\infty,\omega_k\infty)$".) Let $\hat{X} = X \cup \{P_1,\ldots P_n\}$, and let

$$U_k'' = \{P_k\} \cup \{(x , x f_k(x^{-1})) : x^{-1} \in V\}.$$

(Note that $(x, xf_k(x^{-1}))$ does belong to X.) We have a homeomorphism

$$\phi_k'' : U_k'' - \{P_k\} \longrightarrow V - \{o\}$$

which takes $(x,xf_k(x^{-1}))$ to x^{-1}. We <u>define</u> a bijection $\phi_k'' : U_k'' \to V$ by prescribing $\phi_k''(P_k) = 0$. This gives us an atlas for $\hat{X}$, and it is easy to check that the two conditions on page 30 are satisfied, so that $\hat{X}$ acquires a topology which makes it a surface. It is also easy to check that the topology induced on the subset X by the atlas is the same as its topology as a subset of $\mathbb{C}^2$.

Exercises

1. Find an atlas of two charts for the torus, regarded as a surface of revolution in $\mathbb{R}^3$. What is the transition map between the charts?

2. Find a single chart which covers the whole hyperboloid of one sheet $x^2 + y^2 - z^2 = 1$.

3. Prove that stereographic projection $S^2 - \{\text{north pole}\} \to \mathbb{R}^2$ is conformal.

4. (i) Prove that a smooth orientation-preserving homeomorphism f from one open set of R^2 to another is conformal if and only if the matrix of derivatives Df satisfies the Cauchy-Riemann equations, i.e. if and only if f is holomorphic when $\mathbb{R}^2$ is identified with $\mathbb{C}$.

(ii) Deduce the conformality of Mercator's projection from that of stereographic projection.

5. Prove that the chart $(\phi,\theta) \longmapsto (\phi, \sin\theta)$ for the unit sphere (see page 27) is area-preserving.

6. Consider the atlas for the projective plane described in §2, consisting of three charts $\phi_i : U_i \to \mathbb{R}^2$. Find the transition maps between them, and check that they satisfy the conditions (a) and (b) on page 30.

7. Define the Möbius band as the space X of lines in $\mathbb{R}^2$, with the topology given by the atlas of two charts described in §1. What is the transition map? Is it orientation-preserving?

Let $\{\phi_\alpha : U_\alpha \to V_\alpha\}$ be an arbitrary atlas for X. Show that there is a continuous map $(s,t) \longmapsto \gamma_s(t)$ from $[0,\pi] \times [0,2\pi]$ to X such that

(i) for each $s \in [0,\pi]$ the map $t \longmapsto \gamma_s(t)$ is a simple closed curve in X which is completely contained in at least one of the sets U_α, and

(ii) the curves γ_0 and γ_π are the same but described in opposite senses, i.e. $\gamma_0(t) = \gamma_\pi(2\pi-t)$.

Deduce that X is not orientable.

[Take γ_s to be a circle of small radius ε on X with centre at the point (s, 0) in the standard chart. Assume that if $(s,t) \longmapsto \tilde{\gamma}_s(t)$ is a continuous map $[a,b] \times [0, 2\pi] \to \mathbb{R}^2$ such that $\tilde{\gamma}_s$ is a simple closed curve for all $s \in [a,b]$, then $\tilde{\gamma}_a$ and $\tilde{\gamma}_b$ are either both clockwise or both anticlockwise.]

8. If a topology is defined on the set $\Sigma = \mathbb{C} \cup \{\infty\}$ by means of the two charts described on page 29, prove that the resulting space is homeomorphic to the unit sphere S in $\mathbb{R}^3$. [Define a bijection $S \to \Sigma$ by stereographic projection, and prove it and its inverse are continuous.]

9. If a topology is defined on a set X by means of an atlas $\{\phi_\alpha : U_\alpha \to V_\alpha\}$, prove that X is a surface if the conditions (a) and (b) on page 30 are satisfied.
[Show that (a) $\Leftrightarrow$ (X is locally homeomorphic to $\mathbb{R}^2$), and (b) $\Leftrightarrow$ (X is Hausdorff).]

10. Describe the Riemann surface of the function $p(z)^{\frac{1}{2}}$, where p is a polynomial of degree 2n with 2n distinct real roots. Explain in general terms why the complex curve $y^2 = p(x)$, together with its two points at infinity, is a torus with n-1 holes.

§4 Subdivisions and the Euler number

A polyhedron, e.g. a cube or a pyramid, is a solid object bounded by plane faces. Each face is a closed subset of the surface of the polyhedron, and is homeomorphic to a closed disc in the plane. If two faces intersect then they intersect in an edge, which is homeomorphic to the closed unit interval [0,1]. If two edges intersect then they intersect in a single point, called a vertex. If V, E, and F are the number of vertices, edges, and faces of the polyhedron then the number

$$\chi = V - E + F$$

is called the Euler number of the polyhedron. It is well known that for a convex polyhedron $\chi = 2$.

Let us now consider a generalization of this situation. Suppose that X is a closed surface. We shall define an edge on X as the image of any continuous map.

$$f : [0,1] \to X$$

which is 1-1 except that possibly $f(0) = f(1)$. The points $f(0)$ and $f(1)$ are called the ends of the edge.

Suppose we are given a finite set of points of X which we shall call vertices, and also a finite set of edges. We shall say that these constitute a subdivision of X if

(i) each edge begins and ends in a vertex, and passes through no other vertices,

(ii) two edges intersect at most at their ends, and

(iii) if Γ is the union of the edges then each connected

component of X-Γ is homeomorphic to an open disc in $\mathbb{R}^2$.

The closure of a connected component of X-Γ is then called a face.

Examples

The following are examples of subdivisions of a sphere.

(a) 1 vertex at the north pole.
0 edges
1 face

(b) 1 vertex on the equator
1 edge, the equator
2 faces, the hemispheres

(c) 2 vertices at the poles
1 edge, the Greenwich meridian
1 face

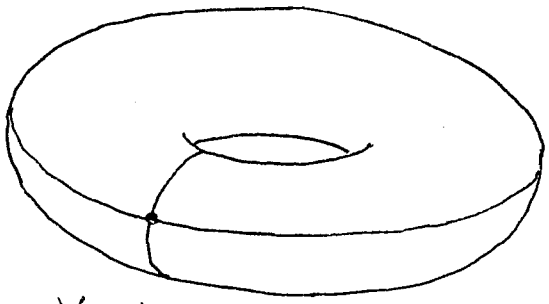
(d) 2 vertices at the poles
2 edges, both meridians
2 faces

(e) the usual subdivision into octants, with 6 vertices, 12 edges, and 8 faces.

Notice that in each case V - E + F = 2.

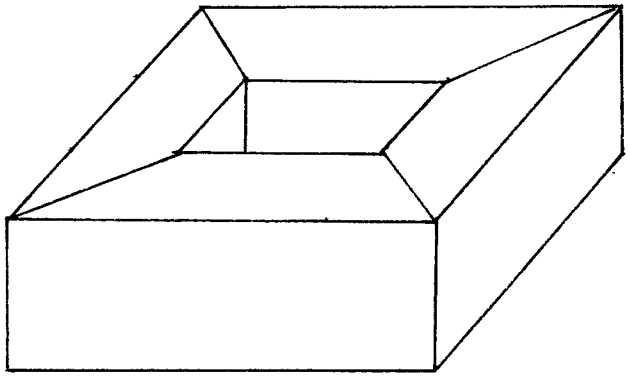
It is a remarkable theorem that for any surface the

Euler number $\chi = V - E + F$ of a subdivision depends only on the surface as a topological space and <u>not</u> on the subdivision. In fact $\chi = 2$ for a sphere, $\chi = 0$ for a torus, $\chi = 2 - 2g$ for a torus with g holes, and $\chi = 2 - k$ for a sphere with k cross-caps.



$V = 1, \quad E = 2, \quad F = 1$

$V - E + F = 0$



$V = 16, \quad E = 32, \quad F = 16$

$V - E + F = 0$

We shall not prove this theorem in full generality in this course, but we shall sketch below an argument which applies to any subdivision of the sphere, and in §10 we shall give a proof which applies to all <u>smooth</u> subdivisions of any surface.

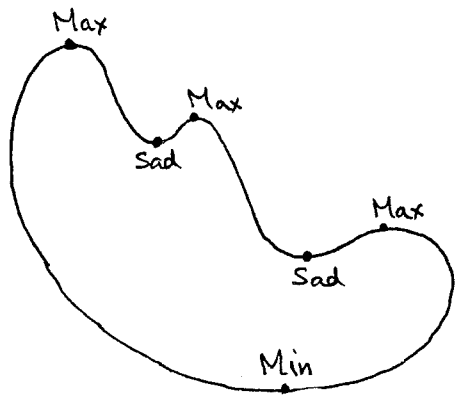
Because the Euler number determines the topological type of a surface (providing we know whether it is orientable or not) it arises in very many situations. We shall mention three here.

(i) Given a smooth function $f : X \to \mathbb{R}$ with isolated non-degenerate critical points (the terminology will be explained in §10) let Max, Min and Sad denote the number of local maxima, local minima, and saddle-points of f. Then

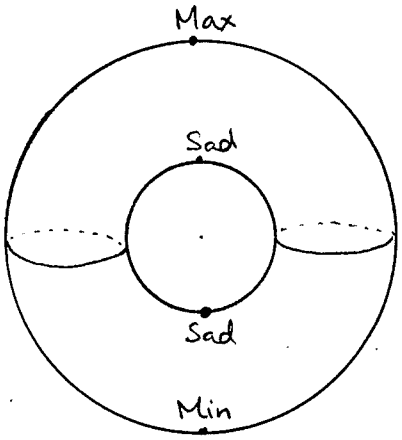
$$\text{Max} - \text{Sad} + \text{Min} = \chi.$$

Examples

In each case the function f is the height above some fixed plane.



Max = 3, Sad = 2, Min = 1
$\chi = 3 - 2 + 1 = 2$



Max = 1, Sad = 2, Min = 1
$\chi = 1 - 2 + 1 = 0$

(ii) Suppose that a fluid is flowing on the surface. Then the number of stationary points of the flow (counted with multiplicities) is χ.

We shall return to these questions in §11.

(iii) "Fermat's last theorem" is the still unproved assertion that if n is an integer greater than two then one cannot find integers x, y, z such that $x^n + y^n = z^n$. This is equivalent to the statement that the algebraic curve $x^n + y^n = 1$ contains no points (x,y) with both coordinates rational. One of the most important mathematical achievements of the last few years has been the theorem of Faltings (1983) which asserts that if f(x,y) is a polynomial with rational coefficients then the algebraic curve f(x,y) = 0 has at most finitely many points with rational coordinates providing the corresponding complex equation f(x,y) = 0 defines a surface for which χ is negative. We shall see presently that when $f(x,y) = x^n + y^n - 1$ we get a surface with $\chi = n(3-n)$. So

Faltings's theorem tells us that when $n > 3$ there are at most finitely many counterexamples to Fermat's theorem.

Sketch of proof that $\chi = 2$ for a sphere

The proof is by induction on the number E of edges in the subdivision. If $E = 0$ then we must have $V = F = 1$, for a sphere with a finite number V of points removed is always connected, and is homeomorphic to an open disc only if $V = 1$. The inductive step is carried out by observing that given any subdivision we can reduce E by one of the following steps:

(i) if there is a vertex contained in only one edge, remove it and the edge;

(ii) if there is an edge contained in a closed cycle of edges, remove it.

In both cases the simplification does not change $V - E + F$. In filling in the details of this argument it will be found that the essential ingredient is the Jordan curve theorem, which asserts that the complement of a simple closed curve on the sphere has exactly two connected components. A complete proof, however, is quite long and difficult.

We should mention that the relation $V - E + F = 2$ for a subdivision of the sphere is a basic tool in graph theory. As an example of its use, let us prove

Theorem (4.1) Given five points in a plane it is impossible to connect each pair by paths which do not cross.

Proof: We may as well replace the plane by a sphere. If we could connect the points we should have a subdivision with $V = 5$ and $E = 10$. Each face would be bounded by at least three edges, and each edge would belong to exactly two faces. Hence $2E \geqslant 3F$. So $F \leqslant 6$, and $V - E + F \leqslant 1$, a contradiction.

The complex curve $x^n + y^n = 1$

We return to the surface $\hat{X}$ defined by the equation $x^n + y^n = 1$ over the complex numbers, together with its n points P_k at infinity. (Cf. the end of §3.) We shall prove that its Euler number is $n(3-n)$. It is orientable (Ex. 5.4), so by Theorem (3.2) it must be a torus with $\frac{1}{2}(n-1)(n-2)$ holes.

The curve $\hat{X} = X \cup \{P_1, \ldots, P_n\}$ maps continuously to the Riemann sphere S by

$$(x,y) \longmapsto x$$
$$P_k \longmapsto \infty .$$

The inverse-image of each point of S, except for the n branch points $B_k = e^{2\pi ik/n}$, consists of exactly n points of $\hat{X}$.

Now let us subdivide the sphere S by taking the points B_k as vertices, and connecting them cyclically to form a polygon. For the subdivision we have $V = n$, $E = n$, $F = 2$. (Notice that $n - n + 2 = 2$.) The inverse-images in $\hat{X}$ of the vertices and edges of the polygon in S provide a subdivision of $\hat{X}$, which has $V = n$ (for there is only one point of $\hat{X}$

above B_k), $E = n^2$, and $F = 2n$. So the Euler number is $n - n^2 + 2n = n(3-n)$.

Exercises

1. Prove that a surface X is connected if and only if it is path-connected, i.e. for every pair of points x, y in X there is a path in X from x to y.

[Write $x \sim y$ if there is a path from x to y. Show that this is an equivalence relation, and consider its equivalence classes.]

2. Let Γ be a connected subset of $\mathbb{R}^2$ which is the union of a finite number E of closed segments of straight lines which intersect only at common end-points. Let V be the total number of end-points. Use the method of the sketch proof on page 42 to give a complete proof that the number of connected components of $\mathbb{R}^2 - \Gamma$ is $E - V + 2$.

3. Prove that on a connected surface any two points can be joined by a path which is an <u>injective</u> map.

[Proceed as in Ex. 1.]

4. Let A_1, A_2, A_3; B_1, B_2, B_3 be six points on a sphere. Prove that one cannot find nine paths on the sphere which link A_i to B_j for each i, j and intersect only at their end-points.

5. Use the method employed for $x^n + y^n = 1$ to show that the Riemann surface associated to the curve $y^2 = p(x)$, where p is a polynomial of degree 2n with distinct roots, is a torus with n-1 holes. (Cf. Ex. 3.10.)

§5 Smooth surfaces

Up to this point in the course we have been concerned with topology. Thus in §1 we mentioned the sphere and the cube as examples of surfaces; but topologically they are identical. From now on we shall be studying the more traditional geometrical questions for which the difference between a sphere and a cube is crucial. We must therefore introduce the concept of a smooth surface. There are two different ways of approaching this: we can think either in terms of abstract surfaces or in terms of surfaces contained in $\mathbb{R}^3$. We shall describe both approaches, as in the end both are needed.

Definition (5.1) A smooth surface is a surface together with a smooth atlas, a smooth atlas being one all of whose transition maps are smooth.

We shall regard two smooth atlases for the same surface as equivalent if each transition map from a chart of the one to a chart of the other is smooth.

Note Our terminology is that a map $f : V_1 \to V_2$ from an open set of $\mathbb{R}^n$ to an open set of $\mathbb{R}^m$ is smooth if all its partial derivatives of all orders exist and are continuous. We shall always think of elements of $\mathbb{R}^n$ and $\mathbb{R}^m$ as column vectors, and shall write $Df(v)$, for $v \in V_1$, for the derivative of f, i.e. the $m \times n$ matrix whose i^{th} column is the i^{th} partial derivative $D_i f(v)$ of f at v.

The important thing about a smooth surface X as defined in (5.1) is that we know what we mean by a smooth function $f : X \to \mathbb{R}$. By definition, f is smooth if for each chart $\phi_\alpha : U_\alpha \to V_\alpha$ the composite map $f \circ \phi_\alpha^{-1} : V_\alpha \to \mathbb{R}$ is smooth. We also know what we mean by, say, a smooth curve in X: a map $\gamma : (a,b) \to X$ is smooth if for each chart the map $\phi_\alpha \circ \gamma$ from $\gamma^{-1}(U_\alpha)$ to V_α is smooth. (Similarly, it should be clear how to define a smooth map from one smooth surface to another.)

For smooth surfaces the question of <u>orientability</u> is easier that for topological surfaces in general. In fact a smooth homeomorphism $f : V_1 \to V_2$ from one connected open set of $\mathbb{R}^2$ to another, with inverse $g : V_2 \to V_1$, is orientation-preserving or reversing according as the Jacobian $\det Df(v)$ is positive or negative for all $v \in V_1$. (The Jacobian cannot vanish, because the matrix Df(v) is invertible with inverse Dg(w), where w = f(v).) We shall not prove this theorem. Instead, as we shall only be interested in <u>smooth</u> homeomorphisms from now on, we shall take the positivity of the Jacobian as the <u>definition</u> of an orientation-preserving map.

Now let us turn to concrete surfaces in $\mathbb{R}^3$.

<u>Definition (5.2)</u> A subset X of $\mathbb{R}^3$ is a smooth surface if for each $x \in X$ there is an open neighbourhood W of x in $\mathbb{R}^3$, and a smooth map $f : W \to \mathbb{R}$ such that

(i) $X \cap W = f^{-1}(0)$, and

(ii) $Df(w)$ does not vanish for $w \in X \cap W$.

Evidently the definition needs to be justified by a proof that such a subset X has an (essentially canonical) smooth atlas. That amounts to the implicit function theorem.

Theorem (5.3) Let $f : W \to \mathbb{R}$ be a smooth map, where W is an open set in $\mathbb{R}^3$. Suppose that $w_\circ = (x_\circ, y_\circ, z_\circ) \in W$ is such that $f(w_\circ) = 0$ and $D_3 f(w_\circ) \neq 0$. Then there is a neighbourhood V of $(x_\circ, y_\circ)$ in $\mathbb{R}^2$, and a smooth map $g : V \to \mathbb{R}$, such that

$$f(x, y, g(x,y)) = 0 \quad \text{for all } (x,y) \in V.$$

Furthermore there is a neighbourhood $W_\circ$ of $w_\circ$ in W such that

$$(x,y,z) \in f^{-1}(0) \cap W_\circ \iff z = g(x,y) \text{ for some } (x,y) \in V.$$

Here $D_3 f$ denotes the partial derivative of f with respect to its third variable. Of course there are equivalent versions of Theorem (5.3) with the roles of the variables (x, y, z) permuted.

The proof of the theorem is given in the Appendix.

Using Theorem (5.3), let us show that a subset X of $\mathbb{R}^3$ satisfying the conditions of (5.2) possesses a smooth atlas.

For any $w_\circ \in X$ the definition gives us a neighbourhood W of $w_\circ$ and a map $f : W \to \mathbb{R}$ such that $(Df)(w_\circ) \neq 0$. Then $D_i f(w_\circ) \neq 0$ for i = 1, 2, or 3. Suppose $D_3 f(w_\circ) \neq 0$, and choose V, g, and $W_\circ$ as in Theorem (5.3). Let $U = X \cap W_\circ$,

and define $\phi : U \to V$ by $\phi(x, y, z) = (x, y)$. Then $\phi : U \to V$ is a homeomorphism, with inverse given by $\phi^{-1}(x, y) = (x, y, g(x,y))$.

The charts defined in this way clearly form an atlas for X, and it is smooth, for if $\tilde{\phi} : \tilde{U} \to \tilde{V}$ is another such chart (got by projecting on to one of the three coordinate planes) then

$$\tilde{\phi} \circ \phi^{-1}(x, y) = (y, g(x,y)) \quad \text{or}$$
$$(x, g(x,y)) \quad \text{or}$$
$$(x, y).$$

In each case the transition map $\tilde{\phi} \circ \phi^{-1}$ is smooth

At this point it is useful to introduce some more terminology. If X is a smooth surface in $\mathbb{R}^3$ we shall say that a chart $\phi : U \to V$ for X is <u>allowable</u> if the inverse map $r = \phi^{-1}$ is a smooth map from V to $\mathbb{R}^3$, <u>and in addition</u> the derivative Dr(v) has rank 2 for all $v \in V$. (I.e. the two vectors $D_1 r(v)$ and $D_2 r(v)$ in $\mathbb{R}^3$ are linearly independent.) We shall call the inverse map r of an allowable chart an <u>allowable parametrization</u>. The charts introduced in the preceding proof were allowable: we had

$$Dr(v) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ D_1 g(v) & D_2 g(v) \end{pmatrix}.$$

The set of the allowable charts is a smooth atlas for X : a proof is given in the appendix. This atlas does not depend on any choices.

The tangent plane

If $\gamma : (a, b) \to \mathbb{R}^3$ is a smooth curve in space, its tangent vector at the point $\gamma(t)$ is $\gamma'(t)$.

Now suppose that X is a smooth surface in $\mathbb{R}^3$, and that $x \in X$.

Definition (5.4) The tangent space to X at x is the set of all tangent vectors at x to all smooth curves in $\mathbb{R}^3$ which pass through x and are contained in X.

The tangent space is in fact a plane, a two-dimensional vector subspace of $\mathbb{R}^3$. That follows from

Theorem (5.5) Let $\phi : U \to V$ be an allowable chart for X such that $x \in U$, and $r : V \to \mathbb{R}^3$ the associated parametrization. Then the tangent space at x is the image space of the linear transformation $Dr(v)$, where $v = \phi(x)$; in other words it is the plane spanned by the vectors $D_1 r(v)$ and $D_2 r(v)$.

Proof: Let $\gamma : (a,b) \to \mathbb{R}^3$ be a curve lying on X such that $\gamma(t) = x$. We may as well suppose that $\gamma((a,b)) \subset U$. Then $\gamma = r \circ \beta$, where $\beta = \phi \circ \gamma : (a,b) \to V$ is a curve in V such that $\beta(t) = v$. By the chain rule the vector $\gamma'(t)$ is the matrix product

$$\gamma'(t) = Dr(v).\beta'(t),$$

i.e. $\gamma'(t) \in$ image $(Dr(v))$.

Conversely, an element of the image of $Dr(v)$ is of

the form $Dr(v).\xi$, where ξ is a vector in $\mathbb{R}^2$. Consider the curve β in V given by

$$t \longmapsto v + t\xi$$

for $|t| < \varepsilon$. Then $\gamma = r \circ \beta$ is a curve in X such that $\gamma(0) = x$, and $\gamma'(0) = Dr(v).\beta'(0) = Dr(v).\xi$.

Normals and orientability

If X is a smooth surface in $\mathbb{R}^3$, and Π_x is its tangent plane at x, then the vectors in $\mathbb{R}^3$ which are perpendicular to Π_x are called normals to X at x. Thus X has two unit normal vectors at each point. If we have an allowable chart $\phi : U \to V$ with $x \in U$ then we can pick out one of the unit normals as the positive one, namely the vector n such that $\{D_1r(v), D_2r(v), n\}$ forms a right-handed frame. (Here $r : V \to \mathbb{R}^3$ is the parametrization inverse to ϕ, and $v = \phi(x)$.) In other words, the positive normal is the unit vector in the direction of the vector product $D_1r(v) \times D_2r(v)$.

Now suppose that x also belongs to another chart $\tilde{\phi} : \tilde{U} \to \tilde{V}$, with $\tilde{\phi}^{-1} = \tilde{r}$ and $\tilde{v} = \tilde{\phi}(x)$. Let the transition map between the charts be $f = \tilde{\phi} \circ \phi^{-1}$. Then $r = \tilde{r} \circ f$, so that by the chain rule

$$D_1r(v) = \alpha\, D_1\tilde{r}(\tilde{v}) + \gamma\, D_2\tilde{r}(\tilde{v})$$

$$D_2r(v) = \beta\, D_1\tilde{r}(\tilde{v}) + \delta\, D_2\tilde{r}(\tilde{v}),$$

where $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = Df(v)$. Thus

$$D_1r(v) \times D_2r(v) = (\alpha\delta - \beta\gamma)\; D_1\tilde{r}(\tilde{v}) \times D_2\tilde{r}(\tilde{v}).$$

Because $\alpha\delta - \beta\gamma = \det Df(v)$, we have proved

Theorem (5.6) Two charts containing x define the same positive normal vector at x if and only if the transition map is orientation-preserving.

Corollary (5.7) If X is an orientable smooth surface in $\mathbb{R}^3$ there is a smooth map $n : X \to \mathbb{R}^3$ such that $n(x)$ is the positive unit normal to X at x.

In particular, X is two-sided.

Definition (5.8) The unit normal map $n : X \to \mathbb{R}^3$ of (5.7) is called the Gauss map of X.

We shall return to the Gauss map in §7.

Conversely, if one can choose a unit normal vector $n(x)$ at each point x of X which varies continuously with x then X is orientable, for we can restrict ourselves to allowable charts such that the frame $(D_1r(v), D_2r(v), n(x))$ is right-handed, and these form an oriented atlas for X. In other words

Theorem (5.9) A smooth surface in $\mathbb{R}^3$ is orientable if and only if it is two-sided.

Exercises

1. Let $\{\phi_\alpha : U_\alpha \to V_\alpha\}$ and $\{\tilde{\phi}_\alpha : \tilde{U}_\alpha \to \tilde{V}_\alpha\}$ be two smooth atlases for a surface X. Let F and $\tilde{F}$ denote the sets of smooth maps $X \to \mathbb{R}$ defined using the first and second

atlases respectively. Prove that $F = \tilde{F}$ if the atlases are equivalent.

[In fact the converse result is also true. Thus a smooth surface is completely described by giving X and F instead of X and an atlas. Even more is true: F is obviously a ring under pointwise addition and multiplication, and the topological space X is completely determined by the ring F alone.]

2. Let X be a smooth surface in $\mathbb{R}^3$. If two different atlases are constructed for X by the method used on page 49, prove that they are equivalent.

3. Let X and Y be smooth surfaces. What should be the definition of a smooth map $f : X \to Y$? If Y is a smooth surface in $\mathbb{R}^3$ prove that $f : X \to Y$ is smooth if and only if $i \circ f : X \to \mathbb{R}^3$ is smooth, where $i : Y \to \mathbb{R}^3$ is the inclusion.

4. Let $f : V_1 \to V_2$ be a homeomorphism between open sets of $\mathbb{R}^2$ which is holomorphic when $\mathbb{R}^2$ is identified with $\mathbb{C}$. Prove that f is orientation-preserving.

Prove that the transition maps of the atlas constructed in §3 for the complex curve $x^n + y^n = 1$, and of the atlas constructed in Ex. 3.10 for the curve $y^2 = p(x)$, are holomorphic, and deduce that these surfaces are orientable.

5. Let X be a smooth surface in $\mathbb{R}^3$, and let $f : X \to \mathbb{R}$ be a smooth function. Show that at each point $x \in X$ there

is a unique tangent vector to X, denoted by $(\mathrm{grad}_X f)(x)$, such that

$$\langle (\mathrm{grad}_X f)(x) , \gamma'(t) \rangle = \frac{d}{dt} \{f(\gamma(t))\}$$

for all smooth curves γ on X such that $\gamma(t) = x$. Deduce that $(\mathrm{grad}_X f)(x) = 0$ if f has a local maximum or minimum at x.

[The uniqueness holds because every vector in the tangent plane Π_x at x is of the form $\gamma'(t)$ for some γ. For the existence, let $r : V \to \mathbb{R}^3$ be an allowable parametrization of X in a neighbourhood of x, and let $g = f \circ r : V \to \mathbb{R}$. Then $\mathrm{grad}_X f = D_1 g . e_1 + D_2 g . e_2$, where $\{e_1, e_2\}$ is the basis of Π_x such that $\langle D_i r , e_j \rangle = \delta_{ij}$.]

6. Let X be a smooth surface in $\mathbb{R}^3$, and let $F : \mathbb{R}^3 \to \mathbb{R}$ be a smooth function. Let $f = F|X$. Prove that $(\mathrm{grad}_X f)(x)$ is the projection of $(\mathrm{grad}\ F)(x)$ on to the tangent plane Π_x to X at x.

Deduce "Lagrange's method of undetermined multipliers", i.e. that statement that $(\mathrm{grad}\ F)(x)$ is normal to X if f has a local maximum or minimum at x. (If $X = g^{-1}(0)$, for some smooth function $g : \mathbb{R}^3 \to \mathbb{R}$, then grad F is normal to X if and only if $\mathrm{grad}\ F = \lambda\ \mathrm{grad}\ g$ for some λ.)

§6 The first fundamental form

For the next five sections we shall be studying the geometry of a smooth surface X in $\mathbb{R}^3$ which is covered by a single chart $\phi: X \to V$, where V is an open set in $\mathbb{R}^2$. Thus $X = r(V)$, where

$$r : V \to \mathbb{R}^3$$

is an injective smooth map such that the vectors $D_1 r(v)$ and $D_2 r(v)$ are linearly independent for each $v \in V$.

For brevity we shall use the phrase "X is a patch of surface" to refer to this situation.

Notation

We shall usually abbreviate the tangent vectors $D_i r(v)$ to $r_i(v)$, and shall usually write just r_i when it is obvious which point of the surface we have in mind. The same convention will also be used with other functions defined in V.

We shall systematically write (u,v) for the coordinates of a point of V : this conflicts, of course, with our previous use of v for a point of V.

Finally, when we are being precise, a point of V is a column vector $\binom{u}{v}$. But for typographical convenience we mostly write it (u,v) all the same.

Lengths of curves

A smooth curve γ on X is a map of the form

$$t \longmapsto \gamma(t) = r(u(t), v(t)),$$

where $t \longmapsto (u(t), v(t))$ is a smooth map $[a,b] \to V$. The length of γ is defined as

$$\mathcal{L}(\gamma) = \int_a^b \|\dot{\gamma}(t)\| \, dt,$$

where $\dot{\gamma}(t)$ denotes d/dt. By the chain rule

$$\dot{\gamma}(t) = \dot{u}(t)\, r_1 + \dot{v}(t)\, r_2,$$

so

$$\mathcal{L}(\gamma) = \int_a^b \{E\dot{u}^2 + 2F\, \dot{u}\dot{v} + G\dot{v}^2\}^{\frac{1}{2}} \, dt , \qquad (6.1)$$

where

$$E = \langle r_1, r_1 \rangle , \; F = \langle r_1, r_2 \rangle \text{ and } G = \langle r_2, r_2 \rangle$$

are three functions $V \to \mathbb{R}$ which in the formula (6.1) are understood to be evaluated at $(u(t), v(t)) \in V$. These functions depend only on the surface X and its parametrization, and not on the particular curve : if we are given E,F,G then we have all the information about the surface which we need to calculate the length of any curve on it.

Definition (6.2) The quadratic form

$$E\,du^2 + 2F\,dudv + G\,dv^2$$

is called the first fundamental form of the surface. (*)

Examples

(i) The unit sphere in $\mathbb{R}^3$ with one meridian removed can be parametrized

$$\begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} \cos u & \cos v \\ \cos u & \sin v \\ \sin u & \end{pmatrix}$$

where u and v are latitude and longitude. Then

$$r_1 = \begin{pmatrix} -\sin u & \cos v \\ -\sin u & \sin v \\ \cos u & \end{pmatrix}, \qquad r_2 = \begin{pmatrix} -\cos u & \sin v \\ \cos u & \cos v \\ o & \end{pmatrix},$$

The first fundamental form is $du^2 + \cos^2 u\, dv^2$.

(ii) The surface of revolution formed by rotating the curve $x = f(z)$ in the XZ plane about the Z-axis can be parametrized

$$\begin{pmatrix} u \\ v \end{pmatrix} \longmapsto \begin{pmatrix} f(u) & \cos v \\ f(u) & \sin v \\ u & \end{pmatrix}.$$

The first fundamental form is

$$(1 + f'(u)^2)\, du^2 + f(u)^2\, dv^2.$$

(*) The first fundamental form is simply a way of writing down the three functions E,F,G and at the same time reminding the reader of the formula (6.1). Thus du and dv are formal symbols which are not meant to have any independent "meaning". What we are really talking about is the quadratic form on the tangent plane to X at the point $r(u,v)$ which to the tangent vector $\xi r_1 + \eta r_2$ assigns its length

$$\|\xi r_1 + \eta r_2\|^2 = E\,\xi^2 + 2F\,\xi\eta + G\,\eta^2.$$

The information contained in the first fundamental form is partly about the surface and partly about the chart. Thus the chart is conformal (i.e. angle-preserving) if and only if $F = 0$ and $E = G$ everywhere (see Ex. (6.4)); and we shall see presently that the chart is area-preserving if $EG - F^2 = 1$. All surfaces possess conformal charts and area-preserving charts. One of our main tasks is to extract from the first fundamental form the information which depends only on the surface and not on the chart.

The first fundamental form does not change if the surface is bent without stretching it. It is useful to introduce the following terminology.

Definition (6.3) Two surfaces X, $\tilde{X}$ in $\mathbb{R}^3$ are isometric if there is a smooth homeomorphism $f : X \to \tilde{X}$ which takes each curve to a curve of the same length.

Then we can state, for the moment without proof, the basic fact about the first fundamental form.

Theorem (6.4) Two smooth patches of surface X and $\tilde{X}$ in $\mathbb{R}^3$ are isometric if and only if they can be parametrized

$$r : V \to \mathbb{R}^3 \quad \text{and} \quad \tilde{r} : V \to \mathbb{R}^3$$

so that they have the same first fundamental form.

Example

The upper half of the cone $x^2 + y^2 = z^2$, slit along the line $x = -z$, $y = 0$, can be parametrized

$$(\sqrt{2}\ u \cos v,\ \sqrt{2}\ u \sin v,\ \sqrt{2}\ u)$$

with $(u,v) \in (0,\infty) \times (-\pi,\pi)$. The first fundamental form is

$$du^2 + 2u^2\, dv^2.$$

This is the same as the first fundamental form of the wedge-shaped piece of $\mathbb{R}^2$, with angle $2\pi/\sqrt{2}$, parametrized

$$(u \cos(\sqrt{2}\, v),\ u \sin(\sqrt{2}\, v)).$$

Theorem (6.4) gives us a way in principle of deciding when surfaces are isometric. But it is not very practical, for it does not help us to decide when we can reparametrize a surface so as to obtain a desired first fundamental form. The most obvious question is when a given patch of surface is isometric to part of a plane, i.e. when it possesses a chart which is exactly to scale. The only surfaces for which this is obviously true are pieces of cylinders and cones. In fact there is another class: the <u>developable surfaces</u> (cf. §1), which are swept out by the tangent line to a curve in space. We shall prove this now by using Theorem (6.4). In §7 we shall prove the more difficult result that no other surfaces besides cones, cylinders, and developables are isometric to the plane.

Suppose that γ is a curve in $\mathbb{R}^3$ parametrized by arc-length.(*) The associated developable surface X can be

---

(*) A curve $u \longmapsto \gamma(u)$ in $\mathbb{R}^3$ is said to be <u>parametrized by arc-length</u> if the length of the curve from $\gamma(0)$ to $\gamma(u)$ is u. The condition for this is clearly that $\|d\gamma/du\| = 1$. It is also obvious that any smooth curve can be parametrized by arc-length.

parametrized with

$$r(u,v) = \gamma(u) + v\dot{\gamma}(u),$$

where $\dot{\gamma}(u) = d\gamma/du$. Then $r_1 = \dot{\gamma} + v\ddot{\gamma}$, and $r_2 = \dot{\gamma}$. We have $\langle\dot{\gamma},\dot{\gamma}\rangle = 1$, so that $\langle\dot{\gamma},\ddot{\gamma}\rangle = 0$. Recall that $\|\ddot{\gamma}\|$ is the curvature κ of γ.

The first fundamental form is

$$(1 + v^2\kappa^2)\, du^2 + 2\, du\, dv + dv^2$$

Now let us choose a plane curve $u \mapsto \rho(u) = (x(u),y(u))$, again parametrized by arc-length, whose curvature is the same function $\kappa(u)$ as for γ. (We can find ρ by solving the system of two linear differential equations

$$\ddot{x} = \kappa(u)\dot{y}$$
$$\ddot{y} = -\kappa(u)\dot{x}.)$$

Then part of $\mathbb{R}^2$ can be parametrized

$$(u,v) \longmapsto \rho(u) + v\dot{\rho}(u),$$

and it will have the same first fundamental form as the developable surface X.

Proof of Theorem (6.4)

The "if" half of the theorem is obvious. Conversely, if X is parametrized by $r : V \to \mathbb{R}^3$, and $f : X \to \tilde{X}$ is a smooth isometry, then we can define $\tilde{r} : V \to \mathbb{R}^3$ by $\tilde{r} = f \circ r$. It follows from the smoothness of f that r is a smooth map (cf. Ex.5.3) so we have

$$\int_a^b \{E\dot{u}^2 + 2F\dot{u}\dot{v} + Gv^2\}^{\frac{1}{2}}\, dt = \int_a^b \{\tilde{E}\dot{u}^2 + 2\tilde{F}\dot{u}\dot{v} + \tilde{G}\dot{v}^2\}^{\frac{1}{2}}\, dt$$

for <u>all</u> smooth curves $(u(t), v(t))$ in V, where $\tilde{E} = \langle \tilde{r}_1, \tilde{r}_1 \rangle$, etc. Let us apply this to the curve given by $u(t) = u_o + t$ and $v(t) = v_o$ for $o \leqslant t \leqslant \varepsilon$, where (u_o, v_o) is some point of V. We find

$$\int_0^\varepsilon E(u_o + t\,,\, v_o)^{\frac{1}{2}} dt = \int_0^\varepsilon \tilde{E}(u_o + t\,,\, v_o)^{\frac{1}{2}} dt.$$

Because

$$\phi(u_o) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_0^\varepsilon \phi(u_o + t)\, dt$$

for any continuous function ϕ, we conclude that $E(u_o, v_o) = \tilde{E}(u_o, v_o)$. Similarly we can show that $G(u_o, v_o) = \tilde{G}(u_o, v_o)$. Then we consider the curve $u(t) = u_o + t$, $v(t) = v_o + t$. The same argument shows that

$$(E + 2F + G)^{\frac{1}{2}} = (\tilde{E} + 2\tilde{F} + \tilde{G})^{\frac{1}{2}}$$

at (u_o, v_o), and hence that $F(u_o, v_o) = \tilde{F}(u_o, v_o)$. To complete the proof of the theorem we need only to show that $\tilde{r}$ is an allowable parametrization, i.e. to show that the vectors $\tilde{r}_1$ and $\tilde{r}_2$ are linearly independent. But that is so because the equations $E = \tilde{E}$, $F = \tilde{F}$, $G = \tilde{G}$ imply that the angle between $\tilde{r}_1$ and $\tilde{r}_2$ is the same as that between r_1 and r_2.

<u>Areas</u>

Apart from the lengths of curves on the surface the first fundamental form also enables us to calculate areas.

Heuristically at least, the infinitesimal parallelogram on X spanned by the tangent vectors $r_1\,du$ and $r_2\,dv$ has area $\|r_1 \times r_2\|\,dudv$. Using the vector identity

$$\langle a \times b,\ a \times b\rangle = \langle a,a\rangle\langle b,b\rangle - \langle a,b\rangle^2$$

we find

$$\|r_1 \times r_2\| = (EG - F^2)^{\frac{1}{2}}. \qquad (6.5)$$

We therefore make

Definition (6.6) The area of the part of $X = r(V)$ corresponding to $V_o \subset V$ is

$$\int_{V_o} (EG - F^2)^{\frac{1}{2}}\,dudv.$$

For this to be a sensible definition we need to know that it does not depend on the parametrization. Suppose that X is described both by $r : V \to \mathbb{R}^3$ and by $\tilde{r} : \tilde{V} \to \mathbb{R}^3$, with corresponding first fundamental forms $E\,du^2 + 2F\,dudv + G\,dv^2$ $\tilde{E}\,d\tilde{u}^2 + 2\tilde{F}\,d\tilde{u}d\tilde{v} + \tilde{G}\,d\tilde{v}^2$. Then there is a smooth bijection $f : V \to \tilde{V}$, which we write

$$(u, v) \longmapsto (\tilde{u}(u,v),\ \tilde{v}(u, v)),$$

such that $\tilde{r}(\tilde{u}, \tilde{v}) = r(u, v)$. We have

$$r_1 = \tilde{u}_1\,\tilde{r}_1 + \tilde{v}_1\,\tilde{r}_2$$
$$r_2 = \tilde{u}_2\,\tilde{r}_1 + \tilde{v}_2\,\tilde{r}_2\ ,$$

so

$$\begin{aligned}(EG - F^2)^{\frac{1}{2}} &= \|r_1 \times r_2\| \\ &= \|(\tilde{u}_1 \tilde{v}_2 - \tilde{u}_2 \tilde{v}_1)(\tilde{r}_1 \times \tilde{r}_2)\| \\ &= |\tilde{u}_1 \tilde{v}_2 - \tilde{u}_2 \tilde{v}_1| \, \|\tilde{r}_1 \times \tilde{r}_2\| \\ &= \det Df \, . \, (\tilde{E}\tilde{G} - \tilde{F}^2)^{\frac{1}{2}} \, .\end{aligned}$$

Thus by the standard theorem about change of variables in multiple integrals we have

$$\int_{V_0} (EG - F^2)^{\frac{1}{2}} \, dudv = \int_{f(V_0)} (\tilde{E}\tilde{G} - \tilde{F}^2)^{\frac{1}{2}} \, d\tilde{u} \, d\tilde{v} \, .$$

The question of realization

It is natural to ask the following question. If we are given an open set V in $\mathbb{R}^2$ and three smooth functions $E, F, G : V \to \mathbb{R}$, can we find a patch of surface $r : V \to \mathbb{R}^3$ whose first fundamental form is $Edu^2 + 2F \, dudv + G \, dv^2$? If so, then

$$E > 0 \, , \quad G > 0 \, , \quad \text{and } EG - F^2 > 0. \tag{6.7}$$

(In other words, the first fundamental form is positive definite.) If these conditions are satisfied the answer to the question is: locally yes, but globally no. Given E, F, G and a point $P \in V$ we can find a neighbourhood V_0 of P and a map $r : V_0 \to \mathbb{R}^3$ which leads to the desired E, F, G in V_0. But usually we cannot extend r to all of V. We shall not prove these statements.

Let us notice, however, that if E, F, G are given satisfying (6.7) then we can assign a "length" $\mathcal{L}(\gamma)$ to every

curve γ in V by the formula (6.1), and then we can define a metric on V by

$$d(P, Q) = \inf \{ \mathcal{L}(\gamma) : \gamma \text{ is a curve in V from P to Q} \}:$$

In §12 we shall study a very important example of a metric defined in this way : the Poincaré metric on the unit disc in $\mathbb{R}^2$. It cannot be realized (except in little patches) as the natural metric of a surface in $\mathbb{R}^3$.

Exercises

1. The catenary is the plane curve with equation $y = \cosh x$. Why is it so called? The catenoid is the surface of revolution obtained by rotating the catenary about the x-axis. Parametrize it and find its first fundamental form.

2. The helicoid is the ruled surface swept out by a straight line which moves like an aeroplane's propeller : the line is always perpendicular to the z-axis, and at time t it passes through the point (0, 0, t) and makes an angle t with the x-axis. Parametrize the helicoid and find its first fundamental form.

3. If one meridian is removed from the catenoid prove that the resulting surface is isometric to part of the helicoid, in such a way that meridians of the catenoid map to rulings on the helicoid. What curve on the helicoid corresponds to the "waist" of the catenoid? [We shall in Ex. 10.4 that there is essentially only one possible isometry between these surfaces.]

4. Two curves on the same patch of surface are given parametrically by $t \longmapsto (u(t), v(t))$ and $t \longmapsto (\tilde{u}(t), \tilde{v}(t))$. Suppose that the curves intersect when $t = 0$, i.e. that $u(0) = \tilde{u}(0)$ and $v(0) = \tilde{v}(0)$. Prove that the angle of intersection θ is given by

$$\cos\theta = \frac{E\,\dot{u}\dot{\tilde{u}} + F(\dot{u}\dot{\tilde{v}} + \dot{v}\dot{\tilde{u}}) + G\,\dot{v}\dot{\tilde{v}}}{\{(E\,\dot{u}^2 + 2F\,\dot{u}\dot{v} + G\,\dot{v}^2)(E\,\dot{\tilde{u}}^2 + 2\,F\,\dot{\tilde{u}}\dot{\tilde{v}} + G\,\dot{\tilde{v}}^2)\}^{\frac{1}{2}}}.$$

Deduce that a chart is conformal if and only if the first fundamental form satisfies $E = G$ and $F = 0$ everywhere.

5. Let $\gamma : [a, b] \to \mathbb{R}^3$ be a curve parametrized by arc-length. Its curvature and torsion at $\gamma(u)$ are denoted by $\kappa(u)$ and $\tau(u)$; we suppose that both are non-vanishing. Let Π_u be the plane through $\gamma(u)$ normal to the curve, and let C_u be the circle in Π_u with centre $\gamma(u)$ and radius ε. Let X be the surface swept out by C_u. Prove that when X is suitably parametrized its first fundamental form is

$$((1-\kappa\varepsilon\cos v)^2 + \varepsilon^2\tau^2)\,du^2 + 2\,\varepsilon^2\tau\,du\,dv + \varepsilon^2\,dv^2 .$$

Prove that the area of X is $2\pi\varepsilon$ times the length of γ. [Why is this question not sensible unless $\varepsilon < \text{Min}\,\kappa(u)^{-1}$?]

6. Notice that the first fundamental form can be defined for a patch of surface given by $r : V \to \mathbb{R}^n$ for any value of n. The torus $\{(z_1, z_2) \in \mathbb{C}^2 : |z_1| = |z_2| = 1\}$ - see Ex. 2.1 - can be parametrized $(u, v) \longmapsto (e^{iu}, e^{iv})$. Identifying $\mathbb{C}^2$ with $\mathbb{R}^4$ in the usual way, prove that the corresponding first fundamental form is $du^2 + dv^2$.

[We shall see in §10 that a closed surface in $\mathbb{R}^3$ can never have the first fundamental form $du^2 + dv^2$.]

§7 The curvature of a surface in $\mathbb{R}^3$

We shall continue to study a patch of surface X in $\mathbb{R}^3$, given by a smooth map $r : V \to \mathbb{R}^3$. Near a point $x = r(u, v)$ the surface is approximated by its tangent plane Π_x at x, which is the plane spanned by the vectors $r_1(u, v)$ and $r_2(u, v)$. The curvature of X at x is the way in which X diverges from Π_x. We use Taylor's series to expand $r(u',v')$ when (u',v') is near (u, v):

$$\begin{aligned} r(u',v') = r(u,v) &+ \{r_1(u,v)\ \Delta u + r_2(u,v)\ \Delta v\} \\ &+ \tfrac{1}{2}\{r_{11}(u,v)\ \Delta u^2 + 2r_{12}(u,v)\Delta u\Delta v + r_{22}(u,v)\Delta v^2\} \\ &+ \{\text{ third-order terms }\}, \end{aligned}$$

where $\Delta u = u'-u$ and $\Delta v = v'-v$. If we neglect the third-order terms then the distance of $r(u',v')$ from the tangent plane Π_x is $\langle n, r(u'v') - r(u,v)\rangle =$

$$\tfrac{1}{2}\{\langle n, r_{11}\rangle\ \Delta u^2 + 2\langle n,r_{12}\rangle\ \Delta u\Delta v + \langle n,r_{22}\rangle\ \Delta v^2 , \quad (7.1)$$

where $n = n(u,v)$ is the positive unit normal to X at x.

The quadratic form (7.1), without the $\frac{1}{2}$, is called the second fundamental form of the surface X. It is traditional to write it as

$$L\ du^2 + 2\ M\ dudv + N\ dv^2,$$

where $L = \langle n,r_{11}\rangle$, $M = \langle n,r_{12}\rangle$, and $N = \langle n,r_{22}\rangle$ are real-valued functions of $(u,v) \in V$.

We shall now show that knowledge of the first and second fundamental forms enables one to calculate the curvature of

any curve on X. (In fact much more is true : the two fundamental forms determine X completely up to a rigid motion of $\mathbb{R}^3$. See Ex. 7.6)

Let $\gamma : (a,b) \to \mathbb{R}^3$ be a curve on X parametrized by arc-length. We shall write $\gamma(t) = r(u(t), v(t))$ as usual. The curvature κ of γ at $\gamma(t)$ is, by definition, the length of the vector $\ddot{\gamma}(t)$. This vector can be decomposed

$$\ddot{\gamma} = \ddot{\gamma}_{tgt} + \ddot{\gamma}_{\perp}$$

into a component $\ddot{\gamma}_{tgt}$ in the tangent plane and a component $\ddot{\gamma}_{\perp}$ normal to the surface. The length of $\ddot{\gamma}_{tgt}$ is called the geodesic curvature κ_g of γ - for we shall see in §8 that $\ddot{\gamma}_{tgt} = 0$ if and only if γ is a geodesic - and the length of $\ddot{\gamma}_{\perp}$ is called the normal curvature κ_n of γ. Thus we have

$$\kappa^2 = \kappa_g^2 + \kappa_n^2 . \qquad (7.2)$$

We give signs to κ_g and κ_n by defining

$$\kappa_g = \langle \ddot{\gamma}, \dot{\gamma} \times n \rangle \quad \text{and} \quad \kappa_n = \langle \ddot{\gamma}, n \rangle .$$

The important thing about κ_n is that at any point it depends only on the direction of γ at that point, i.e. on the unit tangent vector $\dot{\gamma} = \dot{u}r_1 + \dot{v}r_2$.

Theorem (7.3) If γ is a curve on X parametrized by arc-length, then its normal curvature is given by

$$\kappa_n = L\dot{u}^2 + 2M\dot{u}\dot{v} + N\dot{v}^2 .$$

The geodesic curvature, on the other hand, depends on $\dot{u}$, $\dot{v}$, $\ddot{u}$, $\ddot{v}$ and the first fundamental form, but not on L, M, N. (See Ex. 7.5)

Proof of (7.3) By definition $\kappa_n = \langle n, \ddot{\gamma} \rangle$. But

$$\ddot{\gamma} = \frac{d}{dt} (\dot{u}r_1 + \dot{v}r_2)$$

$$= (\ddot{u}r_1 + \ddot{v}r_2) + (\dot{u}^2 r_{11} + 2\dot{u}\dot{v}r_{12} + \dot{v}^2 r_{22}).$$

This gives the desired formula, as $\langle n, r_1 \rangle = \langle n, r_2 \rangle = 0$.

From (7.2) we see that of all the curves on X passing through a point $x = r(u,v)$ in the direction $\xi r_1 + \eta r_2$ the minimal possible curvature κ_n is that of the normal section, i.e. the curve of intersection of X with the plane through x spanned by n and the tangent vector $\xi r_1 + \eta r_2$. Let us now consider how κ_n varies if we rotate the unit vector $\xi r_1 + \eta r_2$ in the tangent plane at a fixed point. In other words we consider the values of $\kappa_n = L\xi^2 + 2M\xi\eta + N\eta^2$ subject to the constraint that $\|\xi r_1 + \eta r_2\|^2 = E\xi^2 + 2F\xi\eta + G\eta^2$ is 1. If we change from $\{r_1, r_2\}$ to an orthonormal basis $\{e_1, e_2\}$ in the tangent plane, then the quadratic form $L\xi^2 + 2M\xi\eta + N\eta^2$ will become, say, $A\xi^2 + 2B\xi\eta + C\eta^2$, while the constraint becomes $\xi^2 + \eta^2 = 1$. We can then rotate the basis $\{e_1, e_2\}$ so that the form $A\xi^2 + 2B\xi\eta + C\eta^2$ becomes diagonal, say

$$\kappa_n = \kappa_1 \xi^2 + \kappa_2 \eta^2.$$

As ξ and η vary subject to $\xi^2 + \eta^2 = 1$ the normal curvature varies between κ_1 and κ_2. If $\kappa_1 \neq \kappa_2$ the eigendirections $\{e_1, e_2\}$ are uniquely determined.

We now introduce some standard terminology.

Definition (7.4)

(i) The extreme values κ_1, κ_2 of the normal curvature at

a point of a surface are called the principal curvatures at that point.

(ii) The directions of the curves with curvatures κ_1, κ_2 are called the principal directions at the point.

(iii) The product $K = \kappa_1\kappa_2$ is called the Gaussian curvature of the surface.

(iv) The average $\frac{1}{2}(\kappa_1 + \kappa_2)$ is called the mean curvature of the surface. (*)

Remarks

(i) The principal directions are not defined at a point where $\kappa_1 = \kappa_2$, but they are perpendicular whenever they are defined. (A point with $\kappa_1 = \kappa_2$ is called an "umbilic".)

(iii) If κ_1 and κ_2 have the same sign - i.e. $K > 0$ - then the surface looks convex or concave, i.e. it stays on one side of its tangent plane. If κ_1 and κ_2 have opposite signs - i.e. $K < 0$ - the surface looks like a saddle, and crosses its tangent plane.

Our next main task is to explain the geometric significance of the Gaussian curvature and the mean curvature. But first let us point out that the change of basis

$$e_1 = \alpha_1 r_1 + \beta_1 r_2$$

$$e_2 = \alpha_2 r_1 + \beta_2 r_2$$

in the tangent plane which converts the fundamental forms $E\xi^2 + 2F\xi\eta + G\eta^2$ and $L\xi^2 + 2M\xi\eta + N\eta^2$ to $\xi^2 + \eta^2$ and $\kappa_1\xi^2 + \kappa_2\eta^2$ can be done in one step. For

(*) Many books define the mean curvature as $\kappa_1 + \kappa_2$.

$$\begin{pmatrix}\alpha_1\\ \beta_1\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}\alpha_2\\ \beta_2\end{pmatrix}$$

are the relative eigenvectors of the two forms, i.e. the vectors such that

$$\begin{pmatrix}L & M\\ M & N\end{pmatrix}\begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix} = \kappa_i \begin{pmatrix}E & F\\ F & G\end{pmatrix}\begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix},$$

normalized so that $E\alpha_i^2 + 2F\alpha_i\beta_i + G\beta_i^2 = 1$. So we have

Theorem (7.5) (i) The principal curvatures κ_1, κ_2 are the roots of the quadratic equation

$$\det\left\{\begin{pmatrix}L & M\\ M & N\end{pmatrix} - \kappa\begin{pmatrix}E & F\\ F & G\end{pmatrix}\right\} = 0.$$

(ii) The Gaussian curvature is $K = (LN-M^2)/(EG-F^2)$.

(iii) The mean curvature is

$$\tfrac{1}{2}\,\frac{LG - 2MF + NE}{EG-F^2}$$

Interpretation of the Gaussian curvature

For a plane curve the curvature is defined as the rate of change of direction per unit length, i.e. as $d\psi/ds$ where ψ is the angle between the tangent and a fixed direction. We can equally well take ψ to be the angle between the normal and a fixed direction. The analogous procedure for a surface is to define the curvature as the rate at which the normal sweeps out solid angle per unit area of the surface. We shall now see that this is the Gaussian curvature.

Recall from §5 that for an oriented surface X in $\mathbb{R}^3$ the Gauss map $n : X \to S$ is the map from X to the unit sphere in $\mathbb{R}^3$ which assigns to $x \in X$ the unit normal vector $n(x)$ at x. Let U be a small neighbourhood of x in X. We shall consider the limit as U contracts to x of

$$\frac{\text{area } n(U)}{\text{area } U}, \tag{7.6}$$

where the area of the piece $n(U)$ of S corresponding to U is taken to be positive or negative according as the map n preserves or reverses orientation.

Theorem (7.7) The limit (7.6) exists, and is the Gaussian curvature at x.

Proof: Suppose that X is parametrized near x by $r(u,v)$ in the usual way, and that $U = r(V)$. The area of U is

$$\int_V \|r_1 \times r_2\| \, dudv = \int_V (EG-F^2)^{\frac{1}{2}} \, dudv.$$

The corresponding area on S is $\int_V \|n_1 \times n_2\| \, dudv$. But the vector $n_1 \times n_2$ is in the direction of n, and when we take the sign into account we have

$$\text{area } n(U) = \int_V \langle n, n_1 \times n_2 \rangle \, dudv.$$

Now $\langle n, n_1 \times n_2 \rangle = \dfrac{\langle r_1 \times r_2, n_1 \times n_2 \rangle}{\| r_1 \times r_2 \|}$

$$= \frac{\langle r_1, n_1 \rangle \langle r_2, n_2 \rangle - \langle r_1, n_2 \rangle \langle r_2, n_1 \rangle}{\| r_1 \times r_2 \|}$$

$$= \frac{LN - M^2}{(EG - F^2)^{\frac{1}{2}}},$$

because $\langle r_1, n_1 \rangle = -\langle r_{11}, n \rangle = -L$, etc.

Thus as U shrinks to a point the ratio of areas becomes

$$\frac{LN - M^2}{EG - F^2} = K,$$

as we want.

Corollary (7.8) For a convex closed surface X in $\mathbb{R}^3$ we have

$$\int_X K dA = 4\pi ,$$

where dA is the element of area of X.

Proof: For a convex surface the Gauss map $n : X \to S$ is a bijection, and we have just proved that KdA is the area of $n(dA)$.

In §10 we shall prove that for any closed surface X one has

$$\int_X K dA = 2\pi\chi,$$

where χ is the Euler number of X.

Two other possible definitions of the Gaussian curvature are given in Ex. 10.5.

Flat surfaces

A surface is called flat if its Gaussian curvature vanishes. In §10 we shall prove that a surface is flat if and only if it is locally isometric to a plane. It is a different matter, however, to decide which concrete surfaces in space are flat. Planes, cones, and cylinders are easily seen to be flat, and so are developable ruled surfaces. We shall now show that these are essentially the only ones. This is slightly vague. The most general flat surface is a patchwork of pieces of planes, cones, cylinders, and developables, all meeting each other along straight lines. To avoid such messy statements we shall here prove only

Theorem (7.9) Let X be a flat surface. Then in the neighbourhood of a point x where the mean curvature does not vanish X is a piece of a cone, a cylinder, or a developable.

Proof: At each point in some neighbourhood of x one principal curvature vanishes and the other does not, so there are two well-defined principal directions, corresponding to orthogonal unit tangent vectors e_1, e_2. We shall suppose that e_2 is the direction in which the principal curvature vanishes. We can reparametrize X in a neighbourhood of x so that r_1 and r_2 are parallel to e_1 and e_2 respectively. (See Appendix.) The first and second fundamental forms then reduce to $Edu^2 + Gdv^2$ and Ldu^2.

We shall first show that the curves u = constant are striaght lines, i.e. that X is a ruled surface, and then we

shall show that a ruled surface is flat if and only if it is developable. For the first, it is enough to show that e_2 is constant when u is constant, i.e. that the partial derivative $e_{2,2}$ vanishes.

We begin with the derivatives of the unit normal n. We have

$$n_1 = -E^{-\frac{1}{2}}Le_1 \quad \text{and}$$

$$n_2 = 0,$$

for $\langle n_1, e_1\rangle = \|r_1\|^{-1}\langle n_1, r_1\rangle = -E^{-\frac{1}{2}}\langle n, r_{11}\rangle = -E^{-\frac{1}{2}}L$, etc.

Hence

$$\begin{aligned}\langle e_{2,2}, e_1\rangle &= -E^{-\frac{1}{2}}L\langle e_{2,2}, n_1\rangle \\ &= E^{-\frac{1}{2}}L\langle e_2, n_{12}\rangle \quad (\text{because } \langle e_2, n_1\rangle = 0) \\ &= 0.\end{aligned}$$

Furthermore $\langle e_{2,2}, e_2\rangle = 0$ because e_2 is a unit vector, and $\langle e_{2,2}, n\rangle = -\langle e_2, n_2\rangle = 0$. So $e_{2,2} = 0$, as we want.

Now let us consider the general ruled surface in $\mathbb{R}^3$ swept out by the line through $\gamma(u)$ in the direction of the unit vector $a(u)$, where γ is a curve in $\mathbb{R}^3$ parametrized by arc-length. The surface is parametrized by

$$r(u, v) = \gamma(u) + va(u),$$

so that $r_{11} = \ddot{\gamma} + v\ddot{a}$, $r_{12} = \dot{a}$, and $r_{22} = 0$. Thus in the second fundamental form $N = 0$, and the Gaussian curvature vanishes if and only if $M = 0$, i.e. if and only if $\langle r_1 \times r_2, \dot{a}\rangle = 0$. But

$$\langle r_1 \times r_2 \, , \, \dot{a} \rangle = \langle (\dot{\gamma} + v\dot{a}) \times a \, , \, \dot{a} \rangle$$

$$= \langle \dot{\gamma} \, , \, a \times \dot{a} \rangle .$$

We have seen (Ex. 1.8) that $\langle \dot{\gamma}, a \times \dot{a} \rangle = 0$ is the condition that the surface is developable.

Exercises

1. Find the first and second fundamental forms for the helicoid given parametrically by (u cos v, u sin v, v). Find the principal directions and the principal curvatures at each point.

2. Find the principal directions and the principal curvatures at a point on a surface of revolution in terms of the curvature κ of the generating curve, the distance ρ from the axis, and the angle ϕ between the axis and the tangent line through the point. Prove that the Gaussian curvature is $\kappa \cos \phi / \rho$.

Prove that the isometry found in Ex. 6.3 between the catenoid and the helicoid takes each point to a point where the Gaussian curvature has the same value. [In §10 we shall prove that this is true for <u>any</u> isometry.]

3. A tractrix is the path of a heavy object which begins at the point (0,1) in $\mathbb{R}^2$ and is dragged slowly by a rope of length 1 held by a person who begins at the origin in $\mathbb{R}^2$ and walks with unit speed along the x-axis. Prove that the tractrix can be described parametrically by

$$x = -(\cos u + \log \tan \tfrac{1}{2} u), \quad y = \sin u.$$

The surface obtained by rotating the tractrix about the x-axis is called the tractoid. Prove that its Gaussian curvature is everywhere -1. Prove that the area of the complete surface is 2π.

Prove that when the tractrix is parametrized by arc-length the first fundamental form of the tractoid is $du^2 + e^{-2u}\, dv^2$.

4. For a surface parametrized in the usual way, express the six quantities $\langle r_i, r_{jk} \rangle$, where i, j, k = 1, 2, in terms of E, F, G.

5. If $t \longmapsto (u(t), v(t))$ describes a curve parametrized by arc-length on a patch of surface, prove that its geodesic curvature is given by

$$(EG - F^2)^{\frac{1}{2}} (\dddot{u}\dot{v} - \ddot{u}\ddot{v}) + (p\,\dot{u}^3 + q\,\dot{u}^2\dot{v} + r\,\dot{u}\dot{v}^2 + s\,\dot{v}^3),$$

where p, q, r, s can be expressed in terms of E, F, G. Prove that $p = (EG - F^2)^{-\frac{1}{2}} (EF_1 - \frac{1}{2} EE_2 - \frac{1}{2} FE_1)$.

6. For a surface X parametrized in the usual way let P be the (3×3)-matrix-valued function of (u, v) given by $P = (r_1, r_2, n)$. Prove that the function P determines X up to a translation in $\mathbb{R}^3$. Prove also that P satisfies the differential equations

$$\partial P/\partial u = PA, \qquad \partial P/\partial v = PB,$$

where $A = C^{-1}D$ and $B = C^{-1}E$, and

$$C = \begin{pmatrix} E & F & 0 \\ F & G & O \\ 0 & 0 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} a_{111} & a_{121} & -L \\ a_{211} & a_{221} & -M \\ L & M & 0 \end{pmatrix}, \quad E = \begin{pmatrix} a_{112} & a_{122} & -M \\ a_{212} & a_{222} & -N \\ M & N & 0 \end{pmatrix},$$

and $a_{ijk} = \langle r_i, r_{jk} \rangle$. Deduce that X is determined by the first and second fundamental forms up to a rigid motion of $\mathbb{R}^3$.

[To obtain the differential equations, notice that $P^tP = C$. For the last part, observe that each row ξ^t of P satisfies the differential equations $\partial\xi/\partial u = A^t\xi$ and $\partial\xi/\partial v = B^t\xi$. Use the fact that a first order system of linear ordinary differential equations possesses a unique solution with a given initial condition.]

§8 Geodesics

Roughly speaking, a geodesic on a smooth surface X is a smooth curve $\gamma : [a, b] \to X$ such that the length $\ell(\gamma)$ of γ is minimal among all smooth curves on X joining $\gamma(a)$ to $\gamma(b)$. But we shall adopt a weaker definition, partly just for convenience, and partly because we want, for example, to count any segment of a great circle on a sphere as a geodesic, even one which goes more than half-way round the sphere.

Definition (8.1) A smooth curve $\gamma : [a,b] \to X$ is a geodesic if for every family $\{\gamma_s\}$ of smooth curves in X such that $\gamma_0 = \gamma$ and, for all s, $\gamma_s(a) = \gamma(a)$ and $\gamma_s(b) = \gamma(b)$, we have

$$\frac{d}{ds} \ell(\gamma_s)\Big|_{s=0} = 0.$$

Here a "family of smooth curves" means that $\gamma_s : [a,b] \to X$ is a curve for each s in an interval $(-\varepsilon,\varepsilon)$, and that $(s,t) \longmapsto \gamma_s(t)$ is a smooth map from $(-\varepsilon,\varepsilon) \times [a,b]$ to X.

In this section we shall derive the equations for a geodesic by the standard method of the calculus of variations. The equations are intrinsic to the surface, in the sense that they involve only the first fundamental form : obviously the geodesics do not change if one bends the surface. But we shall find that geodesics can also be characterized in a completely different non-intrinsic way. This is intuitively obvious : a geodesic is a curve on X whose direction changes as little as possible, i.e. one such that the derivative of

its unit tangent vector is normal to the surface.

To derive the equations we can assume that X is a parametrized patch of surface, for any segment of a geodesic is a geodesic. Let us assume that γ_s is given parametrically by $t \longmapsto (u(s,t), v(s,t))$, and differentiate the expression (6.1) for $\mathcal{L}(\gamma_s)$ with respect to s.

$$\mathcal{L}(\gamma_s) = \int_a^b R^{\frac{1}{2}}\, dt ,$$

where $R = E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2$, and $\dot{u}$ and $\dot{v}$ denote $\frac{\partial u}{\partial t}$ and $\frac{\partial v}{\partial t}$.

Thus

$$\frac{d}{ds}\, \mathcal{L}(\gamma_s) = \tfrac{1}{2} \int_a^b R^{-\frac{1}{2}}\, \frac{\partial R}{\partial s}\, dt.$$

But

$$\frac{\partial R}{\partial s} = (E_1\, \dot{u}^2 + 2F_1\, \dot{u}\dot{v} + G_1\, \dot{v}^2)\, \frac{\partial u}{\partial s} + (E_2\dot{u}^2 + 2F\, uv + G_2\dot{v})\, \frac{\partial v}{\partial s}$$
$$+ 2\,(E\dot{u} + F\dot{v})\, \frac{\partial \dot{u}}{\partial s} + 2(F\dot{u} + G\dot{v})\, \frac{\partial \dot{v}}{\partial s} .$$

Integrating by parts, and observing that

$\frac{\partial u}{\partial s} = \frac{\partial v}{\partial s} = 0$ when t = a,b, we find

$$\frac{d\mathcal{L}}{ds}\Big|_{s=o} = \int_a^b \{P\, \frac{\partial u}{\partial s} + Q\, \frac{\partial v}{\partial s}\}\, dt , \qquad (8.2)$$

where

$$P = \tfrac{1}{2}\, R^{-\frac{1}{2}}(E_1\dot{u}^2 + 2F_1\dot{u}\dot{v} + G_1\, \dot{v}^2) - \frac{\partial}{\partial t}\, \{R^{-\frac{1}{2}}(E\dot{u} + F\dot{v})\}$$

and $Q = \tfrac{1}{2}\, R^{-\frac{1}{2}}(E_2\dot{u}^2 + 2F_2\dot{u}\dot{v} + G_2\, \dot{v}^2) - \frac{\partial}{\partial t}\, \{R^{-\frac{1}{2}}(F\dot{u} + G\dot{v})\}$.

In (8.2) the functions P and Q are evaluated at $s = 0$. They are then functions of t which depend only on the original curve γ and not on the family $\{\gamma_s\}$. The curve γ is a geodesic if and only if the derivative given by (8.2) vanishes for <u>every</u> family of curves containing γ. The necessary and sufficient condition for this is that $P = Q = 0$. The sufficiency is obvious; the necessity is proved by the following argument. Suppose, for example, that P does not vanish for $t = t_0$, where $a < t_0 < b$. If $P(t_0) > 0$ then we can find an interval $(t_0 - \delta, t_0 + \delta)$ in which $P(t) > \frac{1}{2} P(t_0) > 0$. Choose a smooth positive-valued function $\phi : [a, b] \to \mathbb{R}$ such that $\phi(t_0) = 1$ but $\phi(t) = 0$ when t is outside the interval $(t_0 - \delta, t_0 + \delta)$. Then consider the family of curves given by

$$u(s, t) = u(t) + s\phi(t)$$

$$v(s, t) = v(t)$$

for $|s| < \varepsilon$. The expression (8.2) takes the value

$$\int_a^b P(t)\, \phi(t)\, dt \geqslant \tfrac{1}{2} P(t_0) \int_a^b \phi(t)\, dt > 0,$$

a contradiction. So $P = 0$; and similarly $Q = 0$.

To put the result in more manageable form we can assume that γ is parametrized by arc-length, so that $R = 1$. (Note that for the preceding argument to work we must not assume that the curves γ_s of the family are parametrized by arc-length for $s \neq 0$, for then $\partial u/\partial s$ and $\partial v/\partial s$ could not be arbitrary functions of t .)

Theorem (8.3) If γ is parametrized by arc-length then it is a geodesic if and only if

$$\frac{d}{dt}(E\dot{u} + F\dot{v}) = \tfrac{1}{2}(E_1 \dot{u}^2 + 2F_1 \dot{u}\dot{v} + G_1 \dot{v}^2) \quad \text{and}$$

$$\frac{d}{dt}(F\dot{u} + G\dot{v}) = \tfrac{1}{2}(E_2 \dot{u}^2 + 2F_2 \dot{u}\dot{v} + G_2 \dot{v}^2).$$

From these equations we can deduce the non-intrinsic description of a geodesic.

Corollary (8.4) A curve on X is a geodesic if and only if the derivative of its unit tangent vector is normal to X at each point.

Proof of (8.4) Let γ be a curve on X. We can assume it is parametrized by arc-length. Its unit tangent vector is $\dot{\gamma} = \dot{u}r_1 + \dot{v}r_2$. Thus $\ddot{\gamma}$ is normal to X if and only if

$$\langle \frac{d}{dt}(\dot{u}r_1 + \dot{v}r_2), r_i \rangle = 0 \quad \text{for} \quad i = 1,2.$$

But these are precisely the equations of Theorem (8.3). For

$$\langle \frac{d}{dt}(\dot{u}r_1 + \dot{v}r_2), r_1 \rangle = \frac{d}{dt}\langle \dot{u}r_1 + \dot{v}r_2, r_1 \rangle - \langle \dot{u}r_1 + \dot{v}r_2, dr_1/dt \rangle$$

$$= \frac{d}{dt}(E\dot{u} + F\dot{v}) - \{\dot{u}^2\langle r_1, r_{11}\rangle + \dot{u}\dot{v}(\langle r_1, r_{12}\rangle + \langle r_2, r_{11}\rangle) + \dot{v}^2\langle r_2, r_{12}\rangle\}$$

$$= \frac{d}{dt}(E\dot{u} + F\dot{v}) - \tfrac{1}{2}(\dot{u}^2 E_1 + 2\dot{u}\dot{v} F_1 + \dot{v}^2 G_1);$$

and similarly for the other equation.

Corollary (8.4) tells us that the geodesics on X are the trajectories of particles which move freely on X, subject to no forces except the constraint of remaining on X. This gives us

a practical way of finding the geodesics on many surfaces. (Cf. especially Ex. (8.3) It also explains why the length of the component $\ddot{\gamma}$ in the tangent plane is called the "geodesic curvature": this component measures the extent to which γ fails to be a geodesic.

Because the determinant $EG-F^2$ is non-zero the equations for a geodesic can be rewritten in the form

$$\ddot{u} = A(u, v; \dot{u}, \dot{v})$$
$$\ddot{v} = B(u, v; \dot{u}, \dot{v}),$$

where A and B are quadratic forms in $\dot{u}$ and $\dot{v}$ whose coefficients are functions of u and v. The theory of ordinary differential equations tells us that if $u(0)$, $v(0)$, $\dot{u}(0)$, and $\dot{v}(0)$ are given then the equations have a solution $(u(t), v(t))$ defined for t in a neighbourhood of 0. In other words, there is always a geodesic passing through a given point in a given direction. In fact if X is complete as a metric space the geodesic can always be extended indefinitely in both directions (i.e. $\gamma(t)$ is defined for all $t \in \mathbb{R}$), but we shall not prove that here.

For the present we shall make only one application of the existence of geodesics. It is to construct a very useful local parametrization of an arbitrary surface X in the neighbourhood of a chosen point x_0. We begin by choosing a

geodesic segment γ, parametrized by arc-length, such that $\gamma(0) = x_o$. Then for all small v we construct a geodesic γ_v, again parametrized by arc-length, orthogonal to γ and such that $\gamma_v(0) = \gamma(v)$. Let $r(u, v) = \gamma_v(u)$. Then $r_1(0,0)$ and $r_2(0,0)$ are orthogonal unit vectors, so r is an allowable parametrization.

Let us calculate the first fundamental form in this parametrization. We have E = 1 because the curves v = constant are parametrized by arc-length; and because these curves are geodesics we find that $F_1 = 0$ from the equations (8.3). Thus F is independent of u. But F = 0 when u = 0 because γ_v meets γ orthogonally. So F = 0 everywhere, and the first fundamental form is

$$du^2 + G(u, v)\ dv^2. \tag{8.5}$$

Example

If we carry out the preceding construction at a point of the equator of the unit sphere, taking γ to be the equator, then u is latitude and v is longitude, and the first fundamental form is

$$du^2 + \cos^2 u\ dv^2.$$

In this section we have only scratched the surface of the theory of geodesics. We have not mentioned the following natural questions.

Is there a geodesic joining any two points of the surface?

When is there more than one?

When is a geodesic the shortest path between two of its points?

What can be said about the existence of closed geodesics on a surface?

When does a pencil of geodesics emanating from one point come to a focus at another?

In fact the theory of geodesics is one of the most beautiful and well worked-out parts of differential geometry.

Exercises

1. What are the geodesics on a cylinder? Verify directly that the principal normal to a geodesic is normal to the cylinder.

[Preferably do not assume that the base of the cylinder is a circle.]

2. Prove that a meridian on a surface of revolution is a geodesic. When is a parallel of latitude a geodesic on such a surface?

3. Prove that along a geodesic γ on a surface of revolution the product $\rho \sin \psi$ is constant, where ρ is the distance of $\gamma(t)$ from the axis of the surface, and ψ is the angle between $\dot{\gamma}(t)$ and the meridian through $\gamma(t)$. This is called Clairault's relation. What does it have to do with the conservation of angular momentum in mechanics?

Prove that on a spheroid (i.e. the curve obtained by rotating an ellipse about one of its axes) every geodesic which is not a meridian remains always between two parallels of latitude.

4. Let X be the hyperboloid of one sheet $x^2 + y^2 - z^2 = 1$, and let $\gamma : \mathbb{R} \to X$ be a geodesic parametrized by arc-length. Let h be the constant value of the "angular momentum" $\rho \sin \psi$ along γ. (See Ex. 3.) Prove that unless γ is either a meridian or the "waist" of X (i.e. the curve $z = 0$) then $\gamma(t)$ is asymptotic to either a meridian or the waist as $t \to +\infty$, and that the latter case occurs if and only if $h = \pm 1$. Prove that γ remains completely in either the top half or the bottom half of the hyperboloid if $|h| > 1$, while if $|h| < 1$ it goes right through.

[Describe γ in cylindrical polar coordinates (ρ, θ, z). Show that $\rho^2 \dot{\theta} = h$. Show that $\dot{z}$ cannot vanish unless $\rho = |h|$. Consider what can be said about ρ and θ if $z \to L$ as $t \to +\infty$.]

5. Let X be an ellipsoid in $\mathbb{R}^3$. If γ is a geodesic on X let $d(t)$ be the length of the diameter of X parallel to $\dot{\gamma}(t)$, and let $p(t)$ be the distance of the tangent plane at $\gamma(t)$ from the centre of the ellipsoid. Prove that the curvature of γ

is $p(t)/d(t)^2$, and that the product $p(t)d(t)$ is independent of t.

6. Let P be a point on a patch of surface X in $\mathbb{R}^3$, and let e be a tangent vector to X at P. Let γ_θ be the geodesic on X parametrized by arc-length such that $\gamma_\theta(0) = P$ and $\dot{\gamma}_\theta(0)$ makes the angle θ with e. Let $r(u,\theta) = \gamma_\theta(u)$. Assume that $(u,\theta) \longmapsto r(u,\theta)$ is an allowable parametrization of X for $0 < u < R$ and $0 < \theta < 2\pi$. (This is called the <u>geodesic polar</u> coordinate system on X at O.) Prove that the corresponding first fundamental form is $du^2 + u^2\, a(u,\theta)^2\, d\theta^2$, where $a : (-R, R) \times \mathbb{R} \to \mathbb{R}$ ia a smooth function such that $a(0,\theta) = 1$ and $a(-u,\theta) = a(u, \theta + \pi)$. Let C_ρ and A_ρ denote the circumference and area of the geodesic circle on X with centre P and radius ρ (i.e. the circle $u = \rho$). Prove that

$$C_\rho = 2\pi\rho\ (1 + \tfrac{1}{2}k\rho^2 + O(\rho^4)) \text{ , and}$$

$$A_\rho = \pi\rho^2\ (1 + \tfrac{1}{4}k\rho^2 + O(\rho^4)),$$

where

$$k = \frac{1}{2\pi}\int_0^{2\pi} a_{11}\,(0,\theta)\, d\theta.$$

(In Ex. 10.5 we shall see that $-6k$ is the Gaussian curvature of X at P.)

[Observe that $(u,\theta) \longmapsto r(u,\theta)$ is actually a smooth map defined for $(u,\theta) \in (-R,R) \times \mathbb{R}$ and satisfying $r(0,\theta) = P$ and $r(-u,\theta) = r(u, \theta + \pi)$. Thus $r_2\,(0,\theta) = 0$. Observe also that $\|r_{21}\,(0,\theta)\| = 1$.]

§9 Mean curvature and minimal surfaces

In this section we shall explain the geometrical significance of the mean curvature. Suppose that a patch of surface is described by $r : V \to \mathbb{R}^3$, where V is an open set of $\mathbb{R}^2$. Let R be a compact region inside V. The area of r(R) is given by

$$A = \int_R \| r_1 \times r_2 \| \, dudv = \int_R (EG-F^2)^{\frac{1}{2}} \, dudv.$$

Now suppose that the surface is moving in such a way that each point travels with unit speed in a direction normal to the surface. We shall prove that the rate of change of A is

$$\dot{A} = 2\int_R C. \, (EG-F^2)^{\frac{1}{2}} \, dudv,$$

where C is the mean curvature. If we consider a very small region R in which the mean curvature is roughly constant, this means that the proportional rate of change of area $A^{-1}\dot{A}$ is roughly C; and letting R contract to a point we have exactly

Theorem (9.1) The mean curvature at a point of a surface is half the rate of change of area, per unit area, at that point when the surface moves perpendicularly to itself with unit speed.

We shall actually prove a slightly more general result. Suppose that we have a one-parameter family of patches of surface, parametrized by $t \in (-\varepsilon,\varepsilon)$. The

surface at time t is given by

$$(u,v) \longmapsto r(u,v:t),$$

where $r = V \times (-\varepsilon,\varepsilon) \to \mathbb{R}^3$ is a smooth map. We shall suppose that the motion is perpendicular to the surface, i.e. that $\dot{r} = D_3 r$ is orthogonal to $r_1 = D_1 r$ and $r_2 = D_2 r$ at each point. Let A(t) denote the area at time t. We have

Theorem (9.2) $$\dot{A}(t) = 2\int_R \| \dot{r} \| .C.(EG-F^2)^{\frac{1}{2}} \, dudv.$$

Here all the quantities in the integral refer to the surface at time t.

Proof: By definition

$$A(t) = \int \| r_1 \times r_2 \| \, dudv$$

$$= \int \langle n, r_1 \times r_2 \rangle \, dudv,$$

so we must show that

$$\frac{\partial}{\partial t} \langle n, r_1 \times r_2 \rangle = \| \dot{r} \| . \| r_1 \times r_2 \| .C.$$

But $$\frac{\partial}{\partial t} \langle n, r_1 \times r_2 \rangle = \langle \dot{n}, r_1 \times r_2 \rangle + \langle n, \dot{r}_1 \times r_2 \rangle + \langle n, r_1 \times \dot{r}_2 \rangle.$$

Now $\langle \dot{n}, r_1 \times r_2 \rangle = \| r_1 \times r_2 \| \langle \dot{n}, n \rangle = 0$ because n is a unit vector, and

$$\| r_1 \times r_2 \| \langle n, \dot{r}_1 \times r_2 \rangle = \langle r_1 \times r_2, \dot{r}_1 \times r_2 \rangle$$

$$= \langle r_1, \dot{r}_1 \rangle \langle r_2, r_2 \rangle - \langle r_2, \dot{r}_1 \rangle \langle r_1, r_2 \rangle.$$

Because $\dot{r} = \|\dot{r}\| n$ we have

$$\dot{r}_1 = \left(\frac{\partial}{\partial u}\|\dot{r}\|\right) n + \|\dot{r}\| n_1,$$

so that $\langle r_1, \dot{r}_1\rangle = \|\dot{r}\| \langle r, n_1\rangle = - L\|\dot{r}\|,$

and $\langle r_2, \dot{r}_1\rangle = \|\dot{r}\| \langle r_2, n_1\rangle = - M \|\dot{r}\| .$

Thus

$$\|r_1 \times r_2\| \langle n, \dot{r}_1 \times r_2\rangle = \|\dot{r}\| \quad (MF-LG),$$

and similarly

$$\|r_1 \times r_2\| \quad \langle n, r_1 \times \dot{r}_2\rangle = \|\dot{r}\| \quad (MF-NE)$$

Putting everything together we have

$$\frac{\partial}{\partial t}\langle n, r_1 \times r_2\rangle = \|\dot{r}\| \quad (2MF - LG - NE)/(EG-F^2)^{\frac{1}{2}}$$
$$= 2\|\dot{r}\| . C . (EG-F^2)^{\frac{1}{2}}$$

by Theorem (7.5).

The importance of Theorem (9.2) is in connection with minimal surfaces. The problem of finding a surface of minimal area with a prescribed boundary curve is called Plateau's problem, and a solution is called a <u>minimal surface</u>. The best example is a soap film spanning a curve of wire. A minimal surface must have the property that its area is stationary to first order when one makes displacements of the surface which vanish at the boundary. Theorem (9.2) tells us that a necessary and sufficient condition for this is that the mean curvature vanishes everywhere. (There is no loss of generality in considering only displacements normal to the surface, for <u>any</u> family of surfaces can be parametrized in such a way that the displacement is normal: one defines the point $r(u,v;t)$ to be the point obtained

from $r(u,v;0)$ by travelling from the surface at time 0 to the surface at time t along a trajectory which is orthogonal to the family of surfaces.)

Alongside soap films one can study soap bubbles, which are films separating regions of space in which the pressure takes different constant values. If we assume that the surface-tension energy possessed by a film is proportional to its area then Theorem (9.2) tells us that the pressure-difference across a film is proportional to its mean curvature, for the change in energy caused by a small displacement is equal to the work done by the pressure. We conclude that a soap bubble is a surface of constant mean curvature.

Exercises

1. Prove that the catenoid (see Ex. 6.1) is a minimal surface, and that it is the only surface of revolution which is a minimal surface.

2. Prove that the helicoid (see Ex. 6.2) is a minimal surface. [In fact it is the only ruled minimal surface: to prove that is a possible but not so easy exercise for the reader.]

3. Suppose that $r : V \to \mathbb{R}^3$ is a conformal parametrization of a patch of surface X (i.e. $E = G$ and $F = 0$: cf. Ex. 6.3) Prove that X is a minimal surface if and only if r is harmonic, i.e. $r_{11} + r_{22} = 0$.

4. If $r : V \to \mathbb{R}^3$ is a conformal parametrization of a minimal surface X prove that r is the real part of a holomorphic map $f : V \to \mathbb{C}^3$ such that $\|df/dz\|^2 = 0$ for all $z \in V$.

[For $\xi = (\xi_1, \xi_2, \xi_3) \in \mathbb{C}^3$ we write $\|\xi\|^2 = \xi_1^2 + \xi_2^2 + \xi_3^2$. Recall that a real-valued harmonic function on V is always the real part of a holomorphic function.]

If X is the catenoid, prove that Im(f) is a parametrization of the helicoid.

5. Conversely, if $f : V \to \mathbb{C}^3$ is a holomorphic map such that $\|f'(z)\|^2 = 0$ for all $z \in V$, prove that Re(f) is a conformal parametrization of a ruled surface.

Deduce from Exercises 1 and 4 a new proof that the helicoid is a minimal surface.

§10 Gauss's "theorema egregium"

In this section we shall prove the fundamental theorem that the Gaussian curvature is an intrinsic property of a surface, i.e. that it can be expressed in terms of the first fundamental form alone, and is therefore invariant under bending. More concretely expressed, however one bends a patch of surface the solid angle swept out by its normals does not change. Gauss called this result a "remarkable theorem" - "theorema egregium" - and the name has remained popular.

We suppose as usual that a patch of surface is defined by $r : V \to \mathbb{R}^3$. Let us choose at each point of the surface an orthonormal basis e_1, e_2 for the tangent plane, in such a way that e_1 and e_2 are smooth maps $V \to \mathbb{R}^3$. One way to choose e_1, e_2 is to apply the Gram-Schmidt procedure to the standard basis r_1, r_2 of the tangent plane: we shall do this in detail presently. Then e_1, e_2, n is an orthonormal basis for $\mathbb{R}^3$, and we can express the partial derivatives $D_j e_i = e_{i,j}$ in terms of it. We write

$$\begin{aligned}
e_{1,1} &= \alpha_1 e_2 + \lambda_1 n \\
e_{1,2} &= \alpha_2 e_2 + \lambda_2 n \\
e_{2,1} &= -\alpha_1 e_1 + \mu_1 n \\
e_{2,2} &= -\alpha_2 e_1 + \mu_2 n,
\end{aligned}$$

where α_1, α_2, λ_1, λ_2, μ_1, μ_2 are all real-valued functions on V.

(We have used the facts that $\langle e_{i,j}, e_i \rangle = 0$ because e_i is a unit vector, and $\langle e_{1,i}, e_2 \rangle = -\langle e_{2,i}, e_1 \rangle$ because

$\langle e_1, e_2 \rangle = 0.$)

The crucial step towards Gauss's theorem is

Lemma (10.1) $\langle e_{1,1}, e_{2,2} \rangle - \langle e_{1,2}, e_{2,1} \rangle$

$= \lambda_1 \mu_2 - \lambda_2 \mu_1$

$= \alpha_{1,2} - \alpha_{2,1}$

$= (LN-M^2)/(EG-F^2)^{\frac{1}{2}}$

Proof: The first equality is immediate from the definitions. For the second, we have

$$\alpha_{1,2} - \alpha_{2,1} = \frac{\partial}{\partial u} \langle e_1, e_{2,2} \rangle - \frac{\partial}{\partial v} \langle e_1, e_{2,1} \rangle$$

$$= \langle e_{1,1}, e_{2,2} \rangle + \langle e_1, e_{2,21} \rangle - \langle e_{1,2}, e_{2,1} \rangle - \langle e_1, e^{2,12} \rangle$$

$$= \langle e_{1,1}, e_{2,2} \rangle - \langle e_{1,2}, e_{2,1} \rangle .$$

To obtain the third equality, recall that we proved on page 7.6 that $(LN-M^2)/(EG-F^2)^{\frac{1}{2}} = \langle n_1 \times n_2, n \rangle$. But $n = e_1 \times e_2$, so

$$\langle n_1 \times n_2, n \rangle = \langle n_1 \times n_2, e_1 \times e_2 \rangle$$

$$= \langle n_1, e_1 \rangle \langle n_2, e_2 \rangle - \langle n_1, e_2 \rangle \langle n_2, e_1 \rangle$$

$$= \langle n, e_{1,1} \rangle \langle n, e_{2,2} \rangle - \langle n, e_{2,1} \rangle \langle n, e_{1,2} \rangle$$

$$= \lambda_1 \mu_2 - \lambda_2 \mu_1 .$$

To deduce from the lemma that the Gaussian curvature $(LN-M^2)/(EG-F^2)$ can be expressed in terms of E, F, G we have only to show that

when the basis $\{e_1,e_2\}$ is suitably chosen the quantities α_1 and α_2 can be expressed in terms of E,F,G. Let us construct $\{e_1,e_2\}$ from $\{r_1,r_2\}$ by the Gram-Schmidt process. Then

$$e_1 = ar_1$$
$$e_2 = br_1 + cr_2,$$

where a, b, c can be expressed in terms of E, F, G. (In fact $a = E^{-\frac{1}{2}}$, $b = -E^{-\frac{1}{2}} F \Delta^{-\frac{1}{2}}$, and $c = E^{\frac{1}{2}}\Delta^{-\frac{1}{2}}$, where $\Delta = EG-F^2$.) So

$$\begin{aligned}\alpha_1 &= \langle e_{1,1}, e_2\rangle \\ &= a\langle r_{11}, e_2\rangle + a_1\langle r_1, e_2\rangle \\ &= ab\langle r_{11}, r_1\rangle + ac\langle r_{11}, r_2\rangle \\ &= \tfrac{1}{2}ab\, E_1 + ac(F_1 - \langle r_1, r_{21}\rangle) \\ &= \tfrac{1}{2}ab\, E_1 + ac(F_1 - \tfrac{1}{2}E_2),\end{aligned}$$

while

$$\begin{aligned}\alpha_2 &= \langle e_{1,2}, e_2\rangle \\ &= a\langle r_{12}, e_2\rangle + a_2\langle r_1, e_2\rangle \\ &= ab\langle r_{12}, r_1\rangle + ac\langle r_{12}, r_2\rangle \\ &= \tfrac{1}{2}ab\, E_2 + \tfrac{1}{2}\, ac\, G_1.\end{aligned}$$

These formulae are messy, but they prove Gauss's theorem.

The formulae are much more manageable when F=0, i.e. when the parameter lines on the surface are orthogonal. In that case $a = E^{-\frac{1}{2}}$, $b = 0$, and $c = G^{-\frac{1}{2}}$, so that we have

$$\alpha_1 = -\tfrac{1}{2}\frac{E_2}{\sqrt{EG}} \quad \text{and} \quad \alpha_2 = \tfrac{1}{2}\frac{G_1}{\sqrt{EG}}.$$

Substituting this in the formula $(\alpha_{1,2} - \alpha_{2,1})(EG)^{-\frac{1}{2}}$ for the Gaussian curvature gives us

Theorem (10.2) If $F = 0$ then the Gaussian curvature is given by

$$K = -\frac{1}{2\sqrt{EG}}\left\{\frac{\partial}{\partial u}\left(\frac{G_1}{\sqrt{EG}}\right) + \frac{\partial}{\partial v}\left(\frac{E_2}{\sqrt{EG}}\right)\right\}.$$

This is still rather complicated. But when the first fundamental form takes the especially nice form $du^2 + G\,dv^2$ - which we discussed in §8 - it becomes much simpler.

Corollary (10.3) If $E = 1$ and $F = 0$ then

$$K = -G^{-\frac{1}{2}}(\partial/\partial u)^2 G^{\frac{1}{2}}.$$

From this last version of the theorem we can deduce some very important geometrical facts about surfaces with constant Gaussian curvature. In §8 we showed that any surface possesses a local parametrization for which the first fundamental form is $du^2 + G\,dv^2$. Indeed we showed a little more, for the parametrization we found had the additional property that the curve $u = 0$ was a geodesic parametrized by arc-length. From Theorem (8.3) we find that this implies that $G(0, v) = 1$ and $G_1(0, v) = 0$ for all v. We can now prove

Theorem (10.4) (i) A surface with Gaussian curvature zero is locally isometric to a plane.

(ii) A surface with constant positive (resp. negative) Gaussian curvature is locally isometric to a sphere (resp. to a tractoid).

Proof: (i) Write $g = G^{\frac{1}{2}}$. If $K = 0$ then $\partial^2 g/\partial u^2 = 0$ from (10.3), so that $g(u, v)$ is of the form $A(v)u + B(v)$. But we have the boundary conditions that $G(0, v) = 1$ and $G_1(0,v) = 0$. So $g(u, v) = 1$, and the first fundamental form is $du^2 + dv^2$, which proves that the surface is locally isometric to a plane.

(ii) In the same way, if $K = a^2 > 0$ then

$$\partial^2 g/\partial u^2 = -a^2 g,$$

so that $g(u, v) = A(v) \cos au + B(v) \sin au$. This time the boundary conditions show that $A(v) = 1$ and $B(v) = 0$, so that the first fundamental form is

$$du^2 + \cos^2 au.dv^2.$$

This is the first fundamental form of a sphere of radius a, with au = latitude and av = longitude.

Finally, if $K = -a^2 < 0$ the same argument leads to

$$du^2 + \cosh^2 au.\ dv^2.$$

This is the first fundamental form of the spool-shaped surface obtained by rotating the curve $(f(u)\ ,\ \cosh au)$, where $u < a^{-1}\sinh^{-1} a^{-1}$, about the x-axis, where

$$f(u) = \int_0^u \{1 - a^2 \sinh^2 at\}^{\frac{1}{2}}\, dt.$$

We shall see in Ex. 12.10 that this surface is locally isometric to the tractoid.

Exercises

1. If the first fundamental form of a surface is $e^{2f}(du^2 + dv^2)$, prove that the Gaussian curvature is $-e^{-2f}(f_{11} + f_{22})$.

2. If the situation of Ex. 1 prove that the second fundamental form satisfies the Mainardi-Codazzi relations

$$L_2 - M_1 = f_2(L + N)$$

$$M_2 - N_1 = -f_1(L + N) .$$

[Take $e_i = e^{-f} r_i$, and express the condition $e_{i,12} = e_{i,21}$.]

3. Prove that no torus in $\mathbb{R}^3$ is isometric to the torus $|z_1| = |z_2| = 1$ in $\mathbb{C}^2$. (See Ex. 6.6.)

4. Let X be the catenoid with one meridian removed, and let Y be the helicoid. (See Ex. 6.3.) Prove that any two isometries $X \to Y$ differ by a rigid screwing movement of the helicoid along itself.

5. Let C_ρ and A_ρ denote the length and the area of the geodesic circle with centre P and radius ρ on a surface X. Prove that the Gaussian curvature K of X at P is equal to each of the following two limits as $\rho \to 0$:

$$\text{(i)} \quad K = -\frac{3}{\pi} \lim (C_\rho - 2\pi\rho)/\rho^3,$$

$$\text{(ii)} \quad K = -\frac{12}{\pi} \lim (A_\rho - \pi\rho^2)/\rho^4.$$

[Use geodesic polar coordinates as in Ex. 8.6.]

§11 The Gauss-Bonnet theorem

The most striking and beautiful theorem about surfaces is the Gauss-Bonnet theorem. In this course, however, we cannot explain its true importance, which is as the prototype of a whole class of theorems which apply in more general higher-dimensional situations.

There are several versions of the theorem. We shall begin with

Theorem (11.1) Let γ be a smooth simple closed curve on a patch of surface X, enclosing a region R. Then

$$\int_{\gamma} \kappa_g \, ds = 2\pi - \int_{R} K dA,$$

where κ_g is the geodesic curvature of γ, ds is the element of arc-length of γ, K is the Gaussian curvature of X, and dA is the element of area of X. The curve γ is supposed to be described anticlockwise.

Examples

(i) If X is the plane then κ_g is the usual curvature $d\Psi/ds$ (where Ψ is the slope of γ), and we have the obvious fact that $\int (d\Psi/ds) ds = 2\pi$.

(ii) On the unit sphere K = 1, so $\int_R KdA$ is just the area of R. If γ is the equator then $\kappa_g = 0$, and the theorem tells us that the area of the northern hemisphere is 2π.

iii) Any simple closed curve on the unit sphere can be regarded as the boundary of either of two regions, but κ_g

changes sign when one changes one's point of view. Applying (11.1) to each region and adding, we find that the area of the sphere is 4π.

Proof of (11.1) Let us recall Green's theorem, which asserts that if $P, Q : V \to \mathbb{R}$ are two smooth functions defined in an open set V of $\mathbb{R}^2$, and γ is a piecewise-smooth simple closed curve in V bounding a region S, then

$$\int_\gamma (P\,du + Q\,dv) = \int_S (Q_1 - P_2)\, du\,dv.$$

Now suppose that X is parametrized by $r : V \to \mathbb{R}^3$, and, as in §10, choose smooth tangent vector fields $e_1, e_2 : V \to \mathbb{R}^3$ such that $\{e_1, e_2\}$ is an orthonormal basis for the tangent space at each point. We shall apply Green's theorem to the line integral

$$I = \int_\beta \langle e_1, \dot{e}_2 \rangle\, ds,$$

where β is the curve in V such that $\gamma = r \circ \beta$. (We can assume that γ is parametrized by arc-length.) Then

$$\dot{e}_2 = \dot{u} e_{2,1} + \dot{v} e_{2,2},$$

so $P = \langle e_1, e_{2,1} \rangle$ and $Q = \langle e_1, e_{2,2} \rangle$, and

$$Q_1 - P_2 = \langle e_{1,1}, e_{2,2} \rangle - \langle e_{1,2}, e_{2,1} \rangle,$$

which, by Lemma (10.1), is $(LN-M^2)/(EG-F^2)^{\frac{1}{2}}$. Thus

$$I = \int_R K\,dA.$$

On the other hand, let $\theta(s)$ be the angle between the unit tangent vector $\dot{\gamma}(s)$ to γ and the unit vector e_1 at the same point $\gamma(s)$. Thus

$$\dot{\gamma} = e_1 \cos\theta + e_2 \sin\theta.$$

Let $\eta = n \times \dot{\gamma}$ be the unit vector in the tangent plane which is perpendicular to $\dot{\gamma}$. Then

$$\eta = -e_1 \sin\theta + e_2 \cos\theta$$

and

$$\ddot{\gamma} = \dot{\theta}\,\eta + \dot{e}_1 \cos\theta + \dot{e}_2 \sin\theta.$$

The geodesic curvature is therefore given by

$$\begin{aligned} \kappa_g &= \langle \eta , \dot{\gamma} \rangle \\ &= \dot{\theta} - \langle e_1, \dot{e}_2 \rangle . \end{aligned} \qquad (11.2)$$

So $I = \int (\dot{\theta} - \kappa_g) ds$, which completes the proof of (11.1) for $\int \dot{\theta} = 2\pi$.

If the curve γ in Theorem (11.1) is only <u>piecewise</u> smooth, i.e. R is a curvilinear polygon, then we can still apply Green's theorem. The only difference is that the function θ has a jump discontinuity at each corner of the polygon, the jump δ_i at the i^{th} corner being the external angle of the polygon there. Instead of $\int \dot{\theta} = 2\pi$ in the preceding proof we have

$$\int \dot{\theta} = 2\pi - \Sigma\delta_i.$$

In terms of the <u>internal</u> angles $\alpha_i = \pi - \delta_i$ of the polygon, this gives us

<u>Theorem (11.3)</u> If γ is the boundary of a smooth curvilinear

polygon with n sides and (internal) angles $\alpha_1,\ldots,\alpha_n$ on a smooth surface, then

$$\Sigma\,\alpha_i \;=\; (n-2)\pi \;+\; \int_R K\,dA \;+\; \int_\gamma \kappa_g\,ds.$$

In particular, if the polygon is bounded by geodesics then $\Sigma\alpha_i$ exceeds $(n-2)\pi$ by $\int K\,dA$. We recall that the sum of the angles of a plane n-gon is $(n-2)\pi$.

Example

The sum of the angles of a spherical triangle exceeds π by the area of the triangle. Thus an octant has three angles of $\pi/2$, and area $\pi/2$.

Let us now consider a closed surface X which is subdivided by smooth curves into curvilinear polygons, in the sense explained in §4. We apply Theorem (11.3) to each polygon, and add the resulting equations. The sum of all the angles of all the polygons is $2\pi V$, where V is the number of vertices, for the angles at any vertex add to 2π. Because each edge belongs to two polygons, the sum of the contributions "$(n-2)\pi$" is $2\pi(E-F)$, where E and F are the numbers of edges and faces respectively. The sum of the contributions of the geodesic curvatures is zero, for each edge occurs twice in opposite senses, and κ_g changes sign if we reverse the direction of the curve. (See remark (iii) below.) As $\chi = V - E + F$ is the Euler number of X, we have proved

Theorem (11.4) If X is a smooth closed surface, then

$$\int_X K dA = 2\pi\chi,$$

where χ is the Euler number of X.

Remarks

(i) We proved the result for a convex surface by a much more obvious argument in §7.

(ii) We are accepting without proof that every smooth surface does possess a suitable subdivision.

(iii) The proof we have given applies only to orientable surfaces. For if the surface is not orientable the contributions to the sum from $\int \kappa_g ds$ do not occur in opposite pairs. But in fact the theorem is true in all cases, as one can see by subdividing the surface by edges which are piecewise geodesics.

(iv) The proof was given for a surface in $\mathbb{R}^3$. The statement, however, only involves the first fundamental form, i.e. the metric of the surface. The theorem is really a statement about an abstract surface with a metric, and the proof we have given, when properly interpreted, applies to that situation.

Flows on a closed surface

Suppose that we are given a tangent vector ξ_x at each point x of a smooth closed surface X in $\mathbb{R}^3$. We can think of ξ_x as the velocity at x of some fluid which is flowing on the surface. A point where ξ_x vanishes is a

stationary point of the flow. It is well known that on a sphere, for example, any flow has at least one stationary point. We shall now prove

Theorem (11.5) If the flow ξ on X has only a finite number of stationary points then the number of stationary points, when they are counted with their appropriate multiplicities, is the Euler number of X.

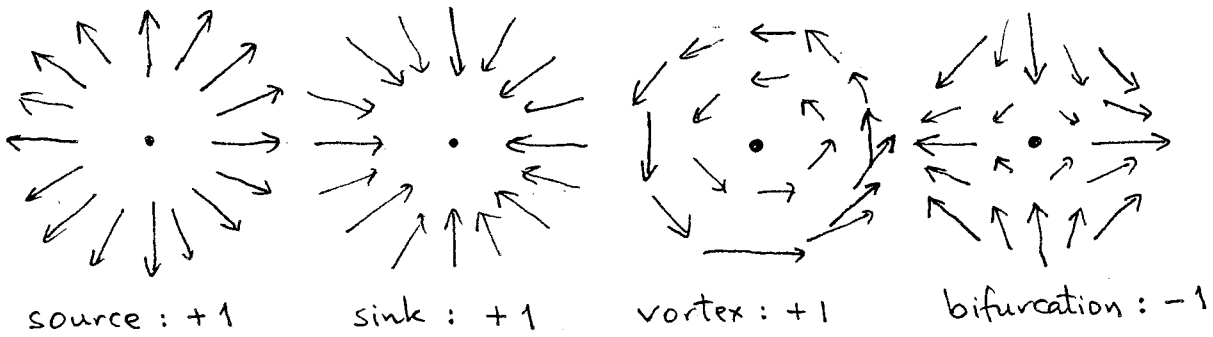
The crucial idea here is the definition of the multiplicity of a stationary point. If $x \in X$ is a stationary point of ξ then we can find a small neighbourhood U of x in X such that $\xi(y) \neq 0$ for $y \in U-\{x\}$. Now let η be another smooth tangent vector field defined, and nowhere-vanishing, in U. (Think of η as providing a reference-direction in U, e.g. $\eta = r_1$ if U is parametrized in the usual way.) Let γ be a small simple closed curve in U which encircles x anticlockwise. Then ξ and η are both non-vanishing on γ, and we define the multiplicity as the winding-number of ξ with respect to η as γ is transversed once, i.e.

$$\text{multiplicity} = \frac{1}{2\pi} \int_\gamma \frac{d\psi}{ds}\, ds,$$

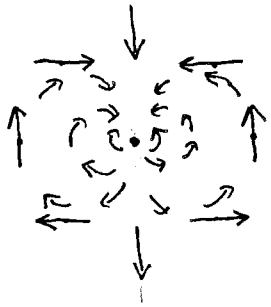
where ψ is the angle between ξ and η at $\gamma(s)$. (Note that although ψ is indeterminate up to multiples of 2π, the derivative $d\psi/ds$ is well-defined.) We leave it to the reader to show that the multiplicity is independent of the choice of η.

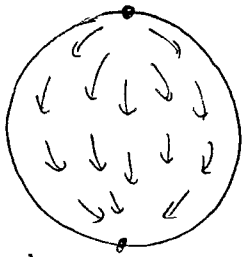
Examples

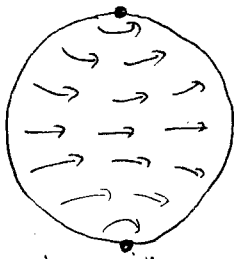
The most common types of stationary points are sources, sinks, vortices, and bifurcations. Their multiplicities are

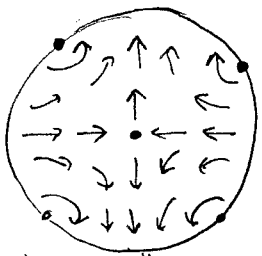


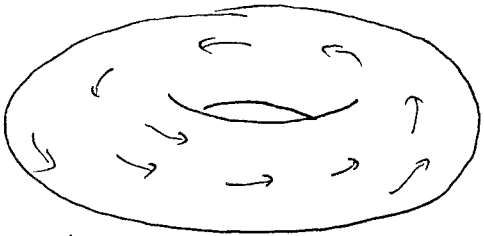
A "dipole-like" flow has a stationary point of multiplicity + 2.

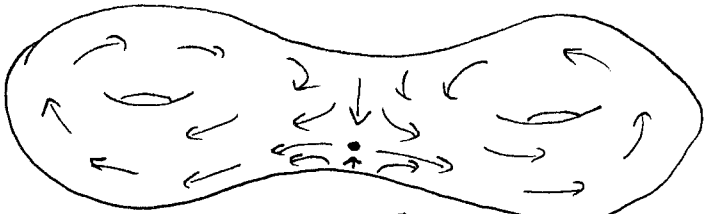



sphere with 1 source 1 sink


sphere with 2 vortices


sphere with 4 vortices 2 bifurcations


torus with no stationary point, $\chi = 0$


2 bifurcation points only, $\chi = -2$.

Proof of Theorem (11.5)

Let $\{x_i\}$ be the stationary points. Choose a small simple closed curve γ_i around each x_i. Let R_i be the small region enclosed by γ_i, and let Y be the part of X outside all the curves γ_i.

At each point $y \in Y$ we can choose an orthonormal basis $\{e_1(y), e_2(y)\}$ for the tangent plane so that $e_1(y)$ is in the direction of $\xi(y)$. Applying the argument of the proof of Theorem (11.1) to the region Y bounded by the curves γ_i gives

$$\int_Y KdA = -\sum_i \int_{\gamma_i} \langle e_1, \dot{e}_2 \rangle \, ds \qquad (11.6)$$

(The minus sign is because the boundary of Y consists of the γ_i oriented clockwise.)

Now let us choose a similar orthonormal basis $\{f_1, f_2\}$ for the tangent planes at the points of the regions R_i. We find

$$\int_{R_i} KdA = \int_{\gamma_i} \langle f_1, \dot{f}_2 \rangle \, ds. \qquad (11.7)$$

Adding (11.6) and (11.7) gives

$$\int_X KdA = \sum_i \int_{\gamma_i} (\langle f_1, \dot{f}_2 \rangle - \langle e_1, \dot{e}_2 \rangle) \, ds.$$

But from (11.2) we have

$$\langle e_1, \dot{e}_2 \rangle = \dot{\theta} - \kappa_g \quad \text{and}$$
$$\langle f_1, \dot{f}_2 \rangle = \dot{\phi} - \kappa_g,$$

where θ and ϕ are the angles between $\dot{\gamma}$ and e_1 and f_1 respectively. Thus

$$2\frac{1}{\pi}\int_X KdA = \sum_i 2\frac{1}{\pi}\int_{\gamma_i}\dot{\psi}ds,$$

where ψ is the angle between e_1 and f_1, i.e. between and f_1. This proves Theorem (11.5), for the left-hand side is the Euler number by (11.4), while the right-hand side is the sum of the multiplicities of the stationary points.

Critical points

Suppose that X is a smooth surface in $\mathbb{R}^3$, and $f : X \to \mathbb{R}$ is a smooth function. We say that f has a critical point at $x \in X$ if the gradient of the composite map $g = f \circ r$ vanishes at v, where $r : V \to \mathbb{R}^3$ is an allowable parametrization of X such that $r(v) = x$. It is easy to check that the definition of a critical point does not depend on the chosen parametrization: in fact x is a critical point of f if and only if the gradient $(\text{grad}_X f)(x)$ vanishes. (See Ex. 5.5 and Ex. 11.6.) Clearly any point at which f has a local maximum or minimum is a critical point.

A critical point is called nondegenerate if the symmetric 2×2 matrix of second derivatives $(D_i D_j g(v))$ is nonsingular. If this matrix is positive-definite or negative-definite then f has a local minimum or maximum If it is nonsingular but indefinite then f has a saddle-point. At points of these three kinds the tangent vector field $\text{grad}_X f$ has a source, a sink, and a bifurcation

respectively. This is intuitively obvious, but we shall not give a detailed proof. If we accept it then we can state

Theorem (11.8) Let $f : X \to \mathbb{R}$ be a smooth function on a closed surface X of Euler number χ. Suppose that f has only a finite number of critical points, all nondegenerate. Then

$$\text{Max} - \text{Sad} + \text{Min} = \chi,$$

where Max, Sad, and Min are the numbers of local maxima, saddle-points, and local minima respectively.

Exercises

1. Calculate

$$\int_\gamma K_g ds \qquad \text{and} \qquad \int_R K dA$$

directly when γ is the boundary of the region R of a surface of revolution bounded by two parallels of latitude. Can you guess a generalization of Theorem (11.1) which applies to an arbitrary region on a surface bounded by a smooth curve?

2. Verify Theorem (11.4) by explicit calculation for the torus in $\mathbb{R}^3$ obtained by rotating the circle $(x-a)^2 + y^2 = b^2$ about the y-axis.

3. Prove that the definition of the multiplicity of a stationary point of a tangent vector field ξ given on page 103 does not depend on the auxiliary vector field η.

[If $\tilde{\eta}$ is another vector field in U, and $\tilde{\psi}$ is the angle between $\tilde{\eta}$ and η, then $d\tilde{\psi}/ds = -(1-f^2)^{-\frac{1}{2}} \dot{f}$, where $f = \cos \tilde{\psi}$. This can be expressed as $P\dot{u} + Q\dot{v}$, where

$P_2 = Q_1$.

4. (i) Draw a diagram of the vector field ξ on $\mathbb{R}^2$ given by $\xi(x,y) = (x^2 - y^2, \quad -2xy)$. What is the multiplicity of the stationary point at the origin?

(ii) Do the same for the vector field $(x^3 - 3xy^2, y^3 - 3x^2y)$.

5. Prove that the definitions of a critical point and of a nondegenerate critical point of a function on a surface do not depend on the chart which is used.

6. Suppose that a patch of surface X is given by $r : V \to \mathbb{R}^3$ in the usual way, and that $r(v) = x$. If $f : X \to \mathbb{R}$ is a smooth function, prove that

$$Dr(v)^* \{(\text{grad}_X f)(x)\} = (\text{grad } g)(v),$$

where $g = f \circ r$, and $Dr(v)^*$ is the adjoint of the isomorphism $Dr(v): \mathbb{R}^2 \to \Pi_x$. (Here Π_x is the tangent plane to X at x.)

Deduce that f has a critical point at x if and only if $(\text{grad}_X f)(x) = 0$.

§ 12 The hyperbolic plane

In this section we shall define a metric, called the Poincaré metric, on the open unit disc in the plane. The resulting metric space is called the hyperbolic plane. Its geometry resembles Euclidean plane geometry, with geodesics playing the role of straight lines. In fact all of Euclid's axioms hold except the so-called "parallel postulate" - the assertion that if a point P is not on a line ℓ there is a unique line through P which does not meet ℓ. The hyperbolic plane was of importance historically, as its discovery ended many centuries of attempts to deduce the parallel postulate from the other axioms, and, more significantly, because it provided the first example of an interesting geometry different from Euclid's.

Euclid's starting point in developing plane geometry was a collection of axioms about the possibility of moving things around. Thus the basis of the definition of length is that the distance AB is equal to the distance A'B' if "when we apply the line AB to the line A'B' so that A falls on A' then the point B falls on B'". In modern language, we assume we are given a group of transformations of the plane which will take any point to any other point and any given line through the first point to a desired line through the second point; and then we prove that the plane possesses a unique metric which is invariant under these transformations.

We shall build up Poincaré's model of the hyperbolic plane in exactly the same way. As our set of points we take the open unit disc $D = \{z \in \mathbb{C} : |z| < 1\}$. We observe

that there is a natural three-parameter group of transformations of D which will take any point to any other and any given direction at the first point to a desired direction at the second. This group is the group G of all holomorphic bijections $f : D \to D$. It is familiar from complex variable theory that for any $a \in D$ the map

$$z \longmapsto \frac{z-a}{1-\bar{a}z} \tag{12.1}$$

is a bijection $D \to D$ which takes a to 0; and the most general element of G which takes a to 0 is got by following (12.1) by a rotation:

$$z \longmapsto e^{i\alpha} \frac{z-a}{1-\bar{a}z},$$

for some $\alpha \in [0,2\pi)$. It is natural, therefore, to look for a metric d on D which is invariant under G, i.e. is such that $d(a,b) = d(f(a),f(b))$ for all $f \in G$.

If such a metric exists, then the distance $d(0,a)$ depends only on $|a|$, for we can take the pair $\{0,a\}$ to $\{0, |a|\}$ by an element of G. Let us write $d(0,a)=p(|a|)$. Then we must have

$$d(a,b) = p\left(\frac{|b-a|}{|1-\bar{a}b|}\right).$$

for any $a, b \in D$, as the map (12.1) takes $\{a, b\}$ to $\{0, (b-a)/1-\bar{a}b)\}$.

We can determine the function p if we add the requirement that distance is to be additive along geodesics. It is reasonable to guess that the real axis

in D will turn out to be a geodesic, so we try to find p so that if a and b are real, with $o<a<b<1$, we have

$$p(a) + p\left(\frac{b-a}{1-ab}\right) = p(b).$$

Differentiating this with respect to b and then putting b=a gives

$$p'(a) = p'(0)/(1-a^2).$$

We can take any value we like for $p'(0)$, for it makes no essential difference if all distances are multiplied by a constant, but the choice $p'(0) = 2$ is traditional, and leads to the simplest formulas. Then

$$p(a) = 2 \tanh^{-1} a.$$

We adopt

Definition (12.2) For $a, b \in D$, let

$$d(a, b) = 2 \tanh^{-1} \left(\frac{|b-a|}{|1-\bar{a}b|} \right).$$

Thus $d(a, b)$ is a positive symmetric function of a and b which vanishes only if $a = b$. To justify the definition we must prove that d satisfies the triangle inequality. Because it was constructed to be invariant under the action of G (see Ex 12.1, 12.2) it is enough to prove

Theorem (12.3) If $a, b \in D$ then

$$d(0, a) + d(0, b) \geqslant d(a, b),$$

with equality if and only if a/b is real and negative.

This, in turn, follows from "the cosine rule for hyperbolic triangles":

Theorem (12.4) If $a, b \in D$ and $\alpha = d(0, a)$, $\beta = d(0, b)$, $\gamma = d(a, b)$, then

$$\cosh \gamma = \cosh \alpha \cosh \beta - \sinh \alpha \sinh \beta \cos \theta,$$

where $\theta = \arg(b/a)$.

This theorem is called the "cosine rule" because when α, β, γ are small, and we use the approximations $\sinh \alpha = \alpha$ $\cosh \alpha = 1 + \frac{1}{2} \alpha^2$, etc., the formula becomes the usual cosine rule for a Euclidean triangle:

$$\gamma^2 = \alpha^2 + \beta^2 - 2\alpha\beta \cos \theta.$$

Furthermore (12.4) implies (12.3) because $\cos \theta \geqslant -1$, so that

$$\begin{aligned} \cosh \gamma &\leqslant \cosh \alpha \cosh \beta + \sinh \alpha \sinh \beta \\ &= \cosh (\alpha + \beta), \end{aligned}$$

with equality only if $\theta = \pi$.

Proof of Theorem (12.4) By the G-invariance of d we can assume that a is real and positive, so that $a = \tanh \frac{1}{2} \alpha$ and $b = e^{i\theta} \tanh \frac{1}{2} \beta$. Then

$$\cosh \alpha = \frac{1 + |a|^2}{1 - |a|^2} \quad \text{and} \quad \cosh \beta = \frac{1 + |b|^2}{1 - |b|^2}.$$

By definition $\tanh \frac{1}{2}\gamma = |b-a|/|1-\bar{a}b|$, so

$$\cosh \gamma = \frac{|1-\bar{a}b|^2 + |b-a|^2}{|1-\bar{a}b|^2 - |b-a|^2}$$

$$= \frac{(1 + |a|^2)(1 + |b|^2) - 2(\bar{a}b + a\bar{b})}{(1-|a|^2)(1 - |b|)}$$

$$= \cosh \alpha \cosh \beta - \sinh \alpha \sin \beta \cos \theta .$$

Geodesics

In any metric space X the length of a continuous curve γ: [a, b] $\to$ X is defined as the supremum of

$$\sum_{i=i}^{n} d(\gamma(a_{i-i}), \gamma(a_i))$$

when $a = a_o < a_1 < \ldots < a_n = b$ runs through all partitions of the interval [a, b]. It follows that γ is a geodesic if its length is equal to $d(\gamma(a), \gamma(b))$.

For the metric defined on D by (12.2) we conclude from (12.3) that any segment of the real axis is a geodesic, and also is the only curve of minimal length joining any two of its points. But we can move any two points of D on to the real axis by an element of G, so we have proved

Theorem (12.5) There is a unique geodesic joining any two points of D.

Let us now recall from complex variable theory that Möbius transformations

(i) take straight lines and circles to straight lines or circles, and

(ii) are conformal, i.e. preserve the angles of intersection of curves.

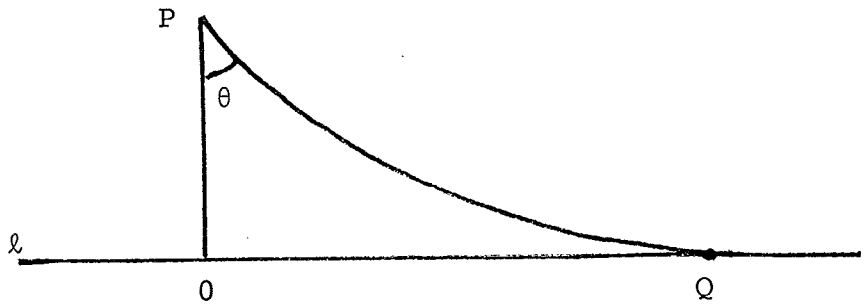
Hence we have

Theorem (12.6) The geodesics in D are the diameters of D and the segments of circles which intersect the boundary of D at right angles.

We have already said that if we take "line" to mean "geodesic" then the geometry of D satisfies all the axioms of Euclidean plane geometry except for the parallel postulate. In particular we can define the angle between two lines, which (because all the isometries of D are conformal maps in the usual sense) turns out to have the Euclidean meaning. (*) To investigate the parallel postulate we consider the following situation.

Let ℓ be a line (i.e. geodesic) in D, and P a point not on ℓ. There is a unique point O on ℓ whose distance a from P is minimal, and the line OP meets ℓ at right angles. (To see this, it is enough to consider the case when ℓ is the real axis and P is on the imaginary axis.)

(*) It is clear from "symmetry" that angles at the centre of D must have their usual values; but any point can be moved to the centre without changing either the hyperbolic or the usual angles.



Let us calculate the angle θ between the lines PO and PQ, where Q is a variable point of ℓ at a distance x from O. By the sine rule (See Ex. 3) for the triangle POQ we have $\sin\theta = \sinh x / \sinh b$, where $b = \alpha(P,Q)$. But $\cosh b = \cosh a \cosh x$ by the cosine rule, so that

$$\sin\theta = \{\cosh^2 a \coth^2 x - \operatorname{cosech}^2 x\}^{-\frac{1}{2}}.$$

As $x \to \infty$ we have $\coth x \to 1$ and $\operatorname{cosech} x \to 0$, so $\sin\theta \to \operatorname{sech} a < 1$. We have proved

Theorem (12.7) A line through P meets ℓ if and only if its angle with PO is less than $\sin^{-1}\operatorname{sech} a$.

The angle $\sin^{-1}\operatorname{sech} a$ is sometimes called the "angle of parallelism at distance a".

Lengths of curves

If two points z and $z + \Delta z$ of D are very close, i.e. if $|\Delta z|$ is very small, then the Poincaré distance $d(z, z + \Delta z)$ is approximately

$$\frac{2|\Delta z|}{1 - |z|^2}.$$

So if γ: [a,b] $\to$ D is a smooth curve the Poincaré length of γ is

$$\mathcal{L}(\gamma) = \int_a^b \frac{2|\dot{\gamma}|}{1-|\gamma|^2}\, dt = \int_a^b \frac{2(\dot{u}^2 + \dot{v}^2)^{\frac{1}{2}}}{1 - u^2 - v^2}\, dt ,$$

where $\gamma(t) = u(t) + iv(t)$. Thus $\mathcal{L}(\gamma)$ is given by our standard formula (6.1), where for the first fundamental form we take

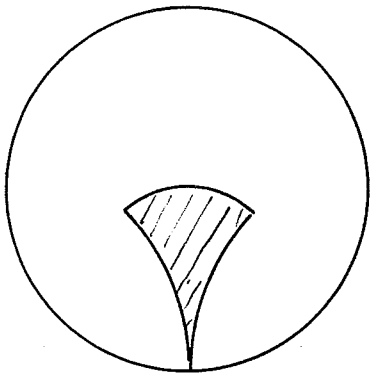
$$\frac{4(du^2 + dv^2)}{(1-u^2 - v^2)^2} \qquad (12.8)$$

This is an abstract first fundamental form which, as far as we know at this point, does not come from any embedding $r : D \to \mathbb{R}^3$ of D as a surface in space. If it <u>did</u> come from a surface in $\mathbb{R}^3$ we could calculate its Gaussian curvature K by Theorem (10.2). In the case when E = G and F = 0 the formula of (10.2) simplifies to

$$K = -\tfrac{1}{2} E^{-1} \Delta(\log E),$$

where Δ is the Laplace operator $(\partial/\partial u)^2 + (\partial/\partial v)^2$. For the form (12.8) we have $E = 4(1-u^2-v^2)^2$, and one readily checks that K = -1.

In §10 we proved that any surface with K = -1 is locally isometric to a tractoid. It can be shown (see Ex. 9) that the shaded region



is isometric to the complete tractoid, except that the boundary curves β and γ become the same meridian on the tractoid. The boundary curve α becomes the cuspidal edge of the tractoid. It can be proved that the whole of D cannot be realized by a surface in $\mathbb{R}^3$: any attempt leads to a surface which "curls up" in some way.

The upper half-plane

The map $z \mapsto i\,\frac{1-z}{1+z}$ is a holomorphic bijection $D \to U$, where U is the half-plane $\{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$. It is often convenient to use this map to identify the hyperbolic plane with U. The geodesics are then the circles orthogonal to the real axis together with all vertical straight lines, and the metric is given by

$$d(a, b) = 2 \tanh^{-1} \frac{|b-a|}{|b-\bar{a}|} .$$

The first fundamental form is

$$\frac{dx^2 + dy^2}{y^2} .$$

The group G of isometries in this realization is the group of all Möbius transformations

$$z \mapsto (az+b)/(cz+d)$$

with a, b, c, d real. We leave the verification of all the preceding facts as exercises.

Areas.

With the first fundamental form (12.8) in mind, we define the hyperbolic area of a region R in D as

$$\frac{4\ dudv}{(1-u^2-v^2)^2}\ .$$

In this section we shall calculate the areas of hyperbolic triangles in D.

As well as triangles proper one can also consider triangles which have one or more vertices at infinity, i.e. on the boundary of D. Such triangles are called asymptotic, biasymptotic, or triasymptotic, according to the number of vertices at infinity.

Any two triasymptotic triangles are congruent, for the group of isometries G will move any three points of the circle $|z| = 1$ to any other two points. Two biasymptotic triangles are congruent if they have the same angle (at their one genuine vertex), for by an isometry we can move any two lines meeting at an angle α to any other two lines meeting at the same angle. Surprisingly enough, the areas of all these infinite "triangles" are finite. We shall prove

Theorem (12.10)

(i) The area of a triasymptotic triangle is π.

(ii) The area of a biasymptotic triangle with angle α is $\pi-\alpha$.

(iii) The area of a triangle with angles α,β,γ is $\pi-\alpha-\beta-\gamma$.

Remark

All three results follow from the Gauss-Bonnet theorem; but we have proved that theorem only for surfaces in $\mathbb{R}^3$.

Proof: We shall prove that (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii), and then we shall prove (i) by direct calculation.

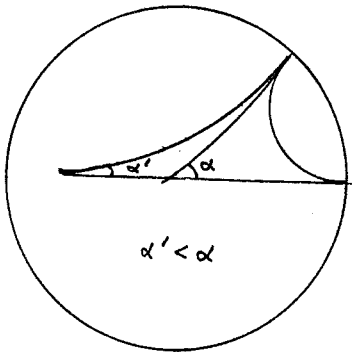


Fig. (a)

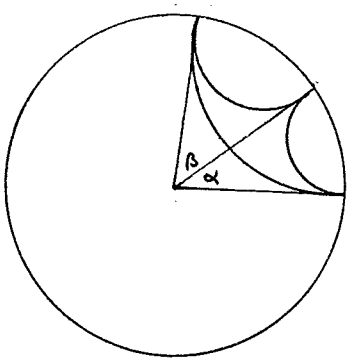


Fig. (b)

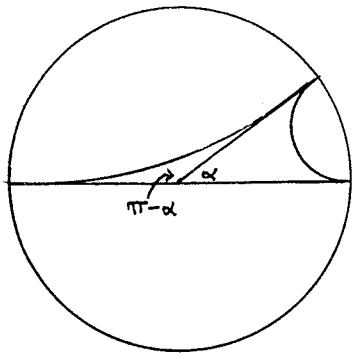


Fig. (c)

Let A_α be the area of a biasymptotic triangle with angle α. From Fig.(a) we see that A_α is a decreasing function of α. From fig.(b) we see that

$$A_\alpha + A_\beta = A_{\alpha+\beta} + \pi,$$

assuming that the area of a triasymptotic triangle is π. If $F(\alpha) = \pi - A_\alpha$ it follows that

$$F(\alpha) + F(\beta) = F(\alpha+\beta).$$

Because F is increasing and additive we conclude that $F(\alpha) = \lambda\alpha$ for some $\lambda > 0$ which does not depend on α, and hence that $A_\alpha = \pi - \lambda\alpha$. But fig. (c) shows that $A_\alpha + A_{\pi-\alpha} = \pi$, and from this it follows that $\lambda = 1$, as desired.

To prove that (ii) $\Rightarrow$ (iii) we consider fig. (d)

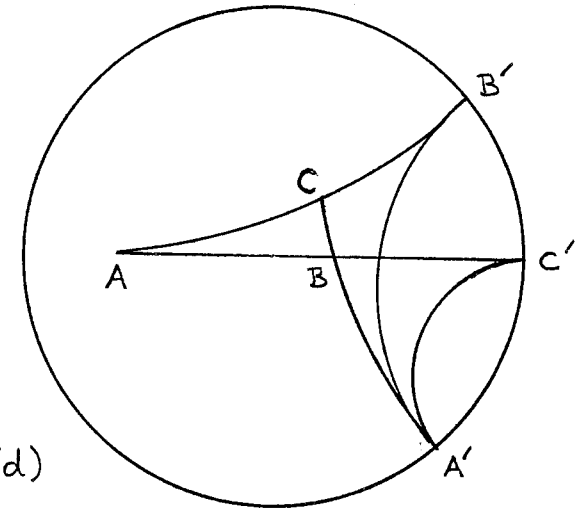


Fig. (d)

We see that

$$\text{area}(ABC) + \text{area}(A'CB') + \text{area}(A'B'C') = \text{area}(AB'C') + \text{area}(A'BC').$$

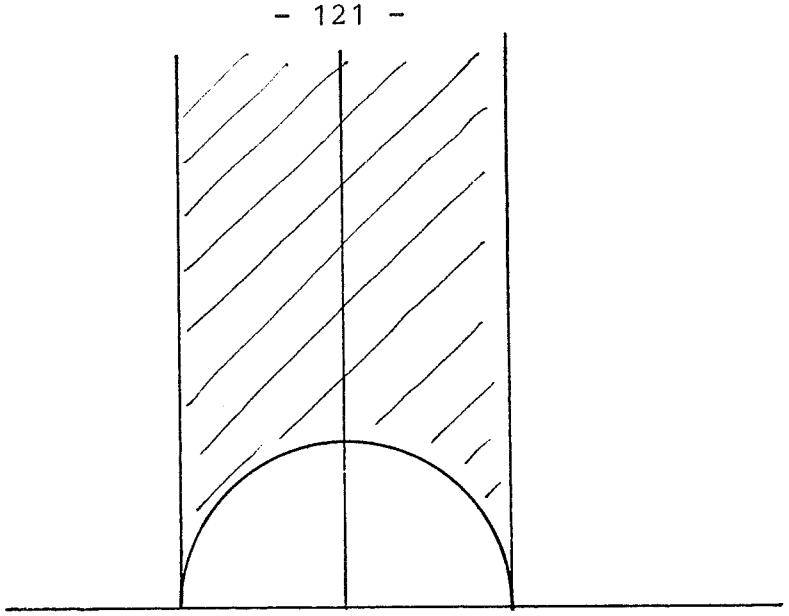
So

$$\text{area}(ABC) + (\pi-(\pi-\gamma)) + \pi = (\pi-\alpha) + (\pi-\beta),$$

and

$$\text{area}(ABC) = \pi - \alpha - \beta - \gamma .$$

Finally, to calculate the area of a triasymptotic triangle it is easiest to work in the upper half-plane. Consider the triangle



bounded by $y = -1$, $y = +1$, and the semicircle $x^2 + y^2 = 1$. From the first fundamental form $(dx^2 + dy^2)/y^2$ we see that the element of area is $dxdy/y^2$. So the triasymptotic triangle has area

$$\int_{-1}^{1} dx \int_{\sqrt{1-x^2}}^{\alpha} \frac{dy}{y^2}$$

$$= \int_{-1}^{1} \frac{dx}{\sqrt{1-x^2}} = \pi .$$

Exercises

1. (a) Prove that any Möbius transformation which preserves the circle $|z| = 1$ and its interior is of the form stated on page 110.

(b) Prove by direct calculation that any such transformation f is an isometry of the Poincaré metric, i.e. that $d(f(a),f(b)) = d(a,b)$ for all $a, b \in D$.

(c) Prove the same result without calculation by using

the facts that (i) the transformations f form a group G, and (ii) the only elements of G which leave 0 fixed are of the form $z \longmapsto e^{i\alpha}z$.

2. Prove that any isometry of the Poincaré metric is either a Möbius transformation or else a Möbius transformation followed by complex conjugation.

[Prove that the only isometries which preserve the origin and also the positive real axis are $z \longmapsto z$ and $z \longmapsto \bar{z}$.]

3. Let ABC be a triangle in the hyperbolic plane which has a right angle at C. Prove the "sine rule"

$$\sin A \,.\, \sinh c \;=\; \sinh a,$$

where $a = d(B,C)$ and $c = d(A,B)$.

[Apply the cosine rule to ABC in two different ways.]

4. Deduce the sine rule

$$\frac{\sin A}{\sinh a} = \frac{\sin B}{\sinh b} = \frac{\sin C}{\sinh c}$$

for an arbitrary hyperbolic triangle from the result of Ex. 3.

5. If α, β, γ are positive, and $\alpha + \beta + \gamma < \pi$, prove that there is a hyperbolic triangle with angles α, β, γ.

6. In the situation of Ex. 3 prove that

$$\sinh^2 a \;+\; \sinh^2 b \;<\; \sinh^2 c\,,$$

and deduce that $A + B < \frac{1}{2}\pi$. Use this to prove that the

sum of the angles of any hyperbolic triangle is less than π.

7. Prove that a hyperbolic circle is simply an ordinary circle in D, but that its hyperbolic centre is usually not its ordinary centre. Prove that a hyperbolic circle of radius a has circumference $2\pi \sinh a$ and area $4\pi \sinh^2 \frac{1}{2}a$.

8. Use Theorem (8.3) to determine the geodesics of the first fundamental form $(dx^2 + dy^2)/y^2$ on the upper half-plane.

9. Let X be the open set $\{x + iy \in \mathbb{C} : -\pi < x < \pi \text{ and } y > 1\}$ of the upper half-plane. Observe that X corresponds to the part of D depicted on page 117. Find a smooth bijection $u : (1,\infty) \longrightarrow (0,\frac{1}{2}\pi)$ such that

$$x + iy \longmapsto$$

$$(-\cos u(y) + \log \cot \tfrac{1}{2}u(y) \ , \ \sin u(y) \cos x \ , \ \sin u(y) \sin x)$$

is an isometry between X with the Poincaré metric and the tractoid described in Ex. 7.3, with one meridian of the tractoid removed.

10. Let X be the open set

$$\{z \in \mathbb{C} : 1 < |z| < e^{2\pi} \text{ and } \tfrac{1}{2}\pi - \alpha < \arg z < \tfrac{1}{2}\pi + \alpha\}$$

of the upper half-plane. Prove that for suitable α the map

$$(u, v) \longmapsto e^v (\tanh u + i \operatorname{sech} u)$$

defines an isometry from the spool-shaped surface described on page 96 (with the parametrization used there) to X with

the Poincaré metric.

[In particular the tractoid is locally isometric to the spool-shaped surface.]

Appendix

In this appendix we shall first recall some definitions and basic facts of differential calculus, then we shall prove the inverse and implicit function theorems, and finally we shall consider four situations in the preceding notes where the theorems were used.

Preliminaries

Let $f : U \to \mathbb{R}^m$ be a map defined in an open set U of $\mathbb{R}^n$. We shall say that f is continuously differentiable if each partial derivative $D_i f(x)$ exists and is continuous for all $x \in U$. The $m \times n$ matrix $Df(x)$ whose i^{th} column is the vector $D_i f(x)$ is called the derivative of f at x.

Let us recall the "chain rule", which asserts that if $f : U \to V$ and $g : V \to \mathbb{R}^k$ are continuously differentiable maps, where U and V are open sets of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, then $g \circ f$ is continuously differentiable, and

$$D(g \circ f)(x) = Dg(f(x)) . Df(x)$$

for all $x \in U$.

The derivative $Df(x)$ is a linear map $\mathbb{R}^n \to \mathbb{R}^m$. Let us recall that the norm of a linear transformation $A : \mathbb{R}^n \to \mathbb{R}^m$ is defined by

$$\|A\| = \sup \{ \|A\,\xi\| : \|\xi\| = 1 \}.$$

Then $\|A\,\xi\| \leqslant \|A\|.\|\xi\|$ for all $\xi \in \mathbb{R}^n$. Notice that $\|A\| \leqslant \sum_{i,j} |A_{ij}|$, where A_{ij} is the $(i,j)^{th}$ entry of A.

The mean value theorem asserts that if f is continuously differentiable then we have

$$\|f(x+h) - f(x)\| \leqslant K\,\|h\|,$$

where $K = \sup \{ \|Df(x+\theta h)\| : 0 \leqslant \theta \leqslant 1\}$. This is proved by applying the single-variable mean value theorem to the function $F : [0, 1] \to \mathbb{R}$ defined by

$$F(t) = \langle u, f(x + th) \rangle,$$

where u is a unit vector parallel to $f(x+h) - f(x)$, and observing that

$$F'(t) = \langle u, Df(x+th).h \rangle,$$

so that $|F'(t)| \leqslant K\,\|h\|$.

The inverse function theorem

We now suppose that $f : U \to \mathbb{R}^n$ is a continuously differentiable map, where U is an open set of $\mathbb{R}^n$. We shall prove that locally f is a bijection providing that the linear transformation Df(x) is invertible. More precisely,

Theorem Suppose that Df(a) is invertible for some $a \in U$. Then there is a neighbourhood V of $b = f(a)$ in $\mathbb{R}^n$, and a continuously differentiable map $g : V \to U$ such that

(i) $g(b) = a$,

(ii) $f(g(y)) = y$ for all $y \in V$,

(iii) $g(V)$ is a neighbourhood of a in $\mathbb{R}^n$.

In particular, f is a bijection in a neighbourhood of a.

Proof: Without loss of generality we can assume that $a = b = 0$, and also that $Df(0) = 1$. (For we can replace f by $x \longmapsto A^{-1}f(x)$, where $A = Df(0)$.) Because Df is continuous we can choose $\varepsilon > 0$ so that $\|Df(x) - 1\| < \frac{1}{2}$ when $\|x\| \leqslant \varepsilon$. Let us define $X = \{ x \in \mathbb{R}^n : \| x \| \leqslant \varepsilon \}$

and $V = \{ y \in \mathbb{R}^n : \| y \| < \frac{1}{2} \varepsilon \}$.

For a fixed $y \in V$ we define $\Phi : U \to \mathbb{R}^n$ by

$$\Phi(x) = x + y - f(x).$$

Notice that $f(x) = y \Leftrightarrow \Phi(x) = x$. Furthermore $\|D\Phi(x)\| \leqslant \frac{1}{2}$ for all $x \in X$, so $\Phi(X) \subset X$.

Define a sequence $\{x_k\}$ in X by $x_o = 0$ and $x_k = \Phi(x_{k-1})$. By the mean value theorem we have

$$\| x_k - x_{k-1} \| \leqslant \tfrac{1}{2} \| x_{k-1} - x_{k-2} \| ,$$

so that $\{ x_k \}$ is a Cauchy sequence. Let its limit be called $g(y) \in X$. Thus g is a map $V \to X$. From $x_k = \Phi(x_{k-1})$ we obtain $g(y) = \Phi(g(y))$, and hence

$$f(g(y)) = y.$$

It is obvious that $g(0) = 0$.

To prove that g is continuously differentiable let us

write $g(y) = x$ and $g(y+k) = x + h$, so that

$$f(x+h) - f(x) = k,$$

and

$$f(x+h) - (x+h) - f(x) + x = k - h.$$

Because $\|Df(x+\theta h) - 1\| < \frac{1}{2}$ this gives

$$\| k - h \| \leqslant \tfrac{1}{2} \|h\|. \qquad (1)$$

We shall now prove that

$$\| g(y+k) - g(y) - Df(x)^{-1}.k\|/\|k\| \to 0 \qquad (2)$$

as $k \to 0$. Applying this when $k = te_i$, where e_i is the i^{th} basis vector of $\mathbb{R}^n$, we find that $D_i g(y)$ exists and is the i^{th} column of $Df(x)^{-1}$. Thus g is continuously differentiable, and $Dg(y) = Df(x)^{-1}$.

To prove (2) we observe

$$g(y+k) - g(y) - Df(x)^{-1}.k = -Df(x)^{-1}.\{f(x+h) - f(x) - Df(x).h\}.$$

So, applying the mean value theorem to the function

$$t \longmapsto f(x + th) - t\, Df(x).h$$

on the interval $[0,1]$ we find

$$\| g(y+k) - g(y) - Df(x)^{-1}.k \| \leqslant \| Df(x)^{-1} \| .\|h\|. R(h),$$

where $R(h) = \sup \{\| Df(x+th) - Df(x)\| : 0 \leqslant t \leqslant 1\}$.

But $R(h) \to 0$ as $h \to 0$ because Df is continuous, and $\|h\| \leqslant 2\|k\|$ from (1), so (2) is proved.

Finally, the restriction of f to X is injective, for if $f(x) = f(x')$ with $x \neq x'$ then

$$(f(x) - x) - (f(x') - x') = -(x-x'),$$

which contradicts the fact that $\| Df - 1 \| < \frac{1}{2}$. It follows that $g(V) = X \cap f^{-1}(V)$, which is a neighbourhood of the origin.

The implicit function theorem

We now suppose that $F : U \to \mathbb{R}^m$ is a continuously differentiable map, where U is an open set of $\mathbb{R}^n$, and $n = k + m > m$. We identify $\mathbb{R}^n$ with $\mathbb{R}^k \times \mathbb{R}^m$. Suppose that $F(a, b) = c$, where $a \in \mathbb{R}^k$ and $b,c \in \mathbb{R}^m$. The implicit function theorem gives a condition under which one can solve the equation $F(x, y) = z$ for y as a function of x and z for (x,z) in a neighbourhood of (a,c).

Theorem In the preceding situation, suppose that the derivative at b of the map $y \longmapsto F(a,y)$ is invertible. Then there is a neighbourhood A of a in $\mathbb{R}^k$, and a neighbourhood C of c in $\mathbb{R}^m$, and a continuously differentiable map $\Phi : A \times C \to \mathbb{R}^m$ such that $\Phi(a,c) = b$, and

$$(x, \Phi(x,z)) \in U \quad \text{for all } (x,z) \in A \times C, \text{ and}$$

$$F(x, \Phi(x,z)) = z.$$

Furthermore there is a neighbourhood W of (a,b) in $\mathbb{R}^{k+m}$ such that if $(x,y) \in W$ and $z \in C$ and $F(x,y) = z$, then $y = \Phi(x,z)$.

Proof: Consider the map $f : U \to \mathbb{R}^n$ defined by

$$f(x, y) = (x, F(x, y)).$$

The derivative $Df(a, b)$ is the $(k + m) \times (k + m)$ matrix

$$\begin{pmatrix} 1 & 0 \\ F_1(a,b) & F_2(a,b) \end{pmatrix},$$

where $F_1(a, b)$ is the derivative of $x \longmapsto F(x, b)$ at a, and $F_2(a,b)$ is the derivative of $y \longmapsto F(a, y)$ at b. The hypotheses imply that $Df(a, b)$ is invertible. By the inverse function theorem we can find a neighbourhood of $f(a, b) = (a, c)$, which we can suppose to be of the form $A \times C$, and a continuously differential map $g : A \times C \to U$, such that $f \circ g$ is the identity. If $g(x, z) = (\theta(x, z), \phi(x, z))$ then $f(g(x, z)) = (\theta(x, z), F(\theta(x, z), \phi(x, z)))$. So $\theta(x, z) = x$, and $F(x, \phi(x, z)) = z$.

Finally, f is injective in a neighbourhood W of (a, b), so in this neighbourhood there can be at most one solution y of $F(x, y) = z$, and it must be $y = \phi(x, z)$.

Applications

1. Complex algebraic curves (See page 10)

We were given a polynomial function of $f : \mathbb{C}^2 \to \mathbb{C}$, and we wanted to solve $f(z, w) = 0$ for w as a function of z. Let us write

$$z = z_1 + iz_2,$$
$$w = w_1 + iw_2,$$
$$f = f_1 + if_2,$$

with z_1, z_2, w_1, w_2, f_1, f_2 all real. The implicit function theorem tells us the condition we used is that the matrix

$$\begin{pmatrix} \partial f_1/\partial w_1 & \partial f_1/\partial w_2 \\ \partial f_2/\partial w_1 & \partial f_2/\partial w_2 \end{pmatrix}$$

is invertible. But the Cauchy-Riemann conditions tell us that $\partial f_1/\partial w_2 = -\partial f_2/\partial w_1$ and $\partial f_2/\partial w_2 = \partial f_1/\partial w_1$. So the determinant of the matrix is

$$(\partial f_1/\partial w_1)^2 + (\partial f_2/\partial w_1)^2 = |\partial f/\partial w|^2.$$

Thus the condition is simply that $\partial f/\partial w \neq 0$, looking just like the real case.

2. Allowable parametrizations (See page 49)

Suppose that $r : V \to \mathbb{R}^3$ and $\tilde{r} : \tilde{V} \to \mathbb{R}^3$ are allowable parametrizations of a smooth surface X in $\mathbb{R}^3$. This means that r and $\tilde{r}$ are smooth maps, and that $Dr(v)$ and $D\tilde{r}(\tilde{v})$ have rank 2 for $v \in V$ and $\tilde{v} \in \tilde{V}$. We wish to prove that the transition map $r^{-1} \circ \tilde{r}$ is smooth in the open subset of $\tilde{V}$ where it is defined: this implies that the allowable charts form an atlas for X.

It is enough to prove that $r^{-1} \circ \tilde{r}$ is smooth in a neighbourhood of each relevant point $\tilde{v} \in \tilde{V}$. Suppose that $x = \tilde{r}(\tilde{v}) = r(v)$. Because the linear transformation $Dr(v) : \mathbb{R}^2 \to \mathbb{R}^3$ has rank 2, the composite $P \circ Dr(v)$ is ibvertible when $P : \mathbb{R}^3 \to \mathbb{R}^2$ is the projection on to one of the coordinate planes. From the inverse function theorem we conclude that the map $P \circ r$ has a smooth inverse $(P \circ r)^{-1}$

defined in a neighbourhood of Px. But then $r^{-1} \circ \tilde{r} = (P\ r)^{-1} \circ P \circ \tilde{r}$ in a neighbourhood of $\tilde{v}$, and so it is smooth.

3. The parametrization of a surface by lines of curvature

(See page 73)

We suppose that at each point of a patch of surface $X = r(V)$ in $\mathbb{R}^3$ there are two well-defined principal directions, corresponding to orthogonal unit tangent vectors $\{e_1, e_2\}$. We wish to reparametrize the surface locally so that r_1 is parallel to e_1 and r_2 to e_2. A curve on X whose tangent vector at each point is in a principal direction is called a line of curvature. It is intuitively obvious that one can find a chart such that the coordinate curves are lines of curvature, but the detailed argument we shall give is surprisingly cumbersome.

Let us first observe that when a smooth tangent vector field $\{\xi(x)\}$ is given on a surface X then one can find a curve $\gamma : (-\varepsilon, \varepsilon) \to X$, with any desired starting point $\gamma(0) = x_0 \in X$, such that $\gamma'(t) = \xi(\gamma(t))$ for all t. For if ξ is expressed in terms of the basis $\{r_1, r_2\}$ by

$$(r(u, v)) = a(u, v)\, r_1(u,v) + b(u, v)\, r_2(u, v),$$

and γ is described parametrically by $(u(t), v(t))$, then finding γ is equivalent to solving the differential equations

$$\dot{u} = a(u, v) \qquad \dot{v} = b(u, v)$$

with a given initial condition $(u(0), v(0))$. Locally such equations can always be solved, and the solution depends

smoothly on $(u(0), v(0) ; t)$.

To construct the desired chart in a neighbourhood of $x_0 = r(u, v)$, first choose a curve α on X such that $\alpha(0) = x_0$ and $\dot{\alpha} = e_1$. Then for each small s choose a curve $t \longmapsto \alpha(s, t)$ such that $\alpha(s,0) = \alpha(s)$ and $\partial\alpha/\partial t = e_2$. Thus $(s, t) \longmapsto \alpha(s, t)$ is a smooth map defined in a neighbourhood of the origin in $\mathbb{R}^2$, and at the origin we have $\alpha_1 = e_1$ and $\alpha_2 = e_2$. Define (u, v) as functions of $(s, t) = r(u, v)$. Then the derivative of $(s, t) \longmapsto (u, v)$ at the origin is

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix},$$

where

$$\begin{aligned} e_1 &= ar_1 + br_2 \\ e_2 &= cr_1 + dr_2 \end{aligned}$$

at x. Because this matrix is invertible we can use the inverse function theorem to express (s, t) in terms of (u, v) locally, and so (s, t) is an allowable parametrization of the surface. It has the property that $\alpha_2 = e_2$ everywhere, but $\alpha_1 \neq e_1$ only when $t = 0$.

Now let us define a curve $s \longmapsto \beta(s, t)$ such that $\beta(0, t) = \alpha(0, t)$ and $\beta_1 = e_1$ everywhere. Just as before we find that $(s, t) \longmapsto \beta(s, t)$ is an allowable parametrization, and we can define a smooth map $(u, v) \longmapsto (\sigma, \tau)$ in a neighbourhood of (u_0, v_0) by $\beta(\sigma, \tau) = r(u, v)$.

The parametrization we want is the one that takes (s, t) to the point of intersection of the curve $\eta \longmapsto \alpha(s, \eta)$ and

and the curve $\xi \longmapsto \beta(\xi, t)$. To obtain it, consider the map $(u, v) \longmapsto (s,\tau)$, where $r(u, v) = \alpha(s, t) = \beta(\sigma,\tau)$. At (u_0, v_0) the derivative of this map is the invertible matrix

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}^{-1},$$

so as usual we can express (u, v) locally in terms of (s, τ) , and $(s, \tau) \longmapsto r(u, v)$ is an allowable parametrization. Furthermore from $r(u, v) = \beta(\sigma, \tau)$ we obtain

$$\partial r/\partial s = (\partial\sigma/\partial s)\beta_1 = (\partial\sigma/\partial s)e_1$$

by regarding σ as a function of s and τ ; and similarly

$$\partial r/\partial\tau = (\partial t/\partial\tau)\alpha_2 = (\partial t/\partial t)e_2.$$

This is what we want.

4. Geodesic polar coordinates (See page 86)

At a point $r(u_0, v_0)$ of a surface let us define a geodesic $(u(t), v(t))$ by solving the equations (8.3) with the initial conditions

$$u(0) = u_0 \qquad v(0) = v_0$$
$$\dot{u}(0) = \xi \qquad \dot{v}(0) = \eta .$$

(The equations (8.3) were derived for a geodesic parametrized by arc-length. They imply. however, that $E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2$ is constant, and so any curve which satisfies them is automatically parametrized proportionally to arc-length, and is a geodesic. Notice also that if $(u(t), v(t))$ is a solution then so is $(u(ct), v(ct))$ for any constant c.)

Having found the geodesic $(u(t), v(t))$, let us regard the point $(u(1), v(1)) = (u(1;\xi,\eta), v(1;\xi,\eta)$ as a function of (ξ, η). The derivative of $(\xi, \eta) \longmapsto (u(1), v(1))$ at the origin is the identity matrix, for

$$u(1; \xi, 0) = u(\xi; 1, 0),$$
$$u(1; 0, \eta) = u(\eta; 0, 1),$$

and similarly for v. Thus $(\xi, \eta) \longmapsto (u(1), v(1))$ is an allowable parametrization when (ξ, η) is in a suitable neighbourhood of the origin.

Geodesic polar coordinates are obtained from the chart just found by choosing an orthonormal basis $\{e_1, e_2\}$ for the tangent plane at $r(u_0, v_0)$ and composing the previous map with the map $(\rho, \theta) \longmapsto (\xi, \eta)$, where

$$\rho.(e_1 \cos\theta + e_2 \sin\theta) = \xi r_1 + \eta r_2.$$