Differential Geometry

J.B. Cooper

1995

Inhaltsverzeichnis

1 CURVES AND SURFACES—INFORMAL DISCUSSION

We begin with an informal discussion of curves and surfaces, concentrating on methods of describing them. We shall illustrate these with examples of classical curves and surfaces which, we hope, will give more content to the material of the following chapters. In these, we will bring a more rigorous approach.

Curves in $\mathbf{R}^2$ are usually specified in one of two ways, the direct or parametric representation and the implicit representation.
For example, straight lines have a
direct representation as

$$\{tx + (1-t)y : t \in \mathbf{R}\}$$

i.e. as the range of the function

$$\phi : t \mapsto tx + (1-t)y$$

(here x and y are **distinct** points on the line) and an
implicit representation:

$$\{(\xi_1, \xi_2) : a\xi_1 + b\xi_2 + c = 0\}$$

(where $a^2 + b^2 \neq 0$) as the **zero set** of the function

$$f(\xi_1, \xi_2) = a\xi_1 + b\xi_2 - c.$$

Similarly, the unit circle has a direct representation

$$\{(\cos t, \sin t) : t \in [0, 2\pi[\}$$

as the range of the function $t \mapsto (\cos t, \sin t)$ and an implicit representation $\{x : \xi_1^2 + \xi_2^2 = 1\}$ as the set of zeros of the function $f(x) = \xi_1^2 + \xi_2^2 - 1$.

We see from these examples that the direct representation displays the curve as the image of a suitable function from $\mathbf{R}$ (or a subset thereof, usually an interval) into two dimensional space, $\mathbf{R}^2$. A good model for this is to regard the independent variable t as time and the curve as the path covered by a moving particle. The implicit definition specifies the curve as the set $\{x : f(x) = 0\}$ of zeros of a function of two variables. These can be conveniently grasped intuitively as one of the contours $f = c$ of the surface of the form $\xi_3 = f(\xi_1, \xi_2)$ (figure 1).

Examples of curves with implicit representations are
the ellipse

$$\frac{\xi_1^2}{a^2} + \frac{\xi_2^2}{b^2} = 1$$

the parabola

$$\xi_2 - \xi_1^2 = 0.$$

Certain of the classical descriptions of curves as the loci of points submitted to certain constraints can be conveniently interpreted in this way. For example, an ellipse is often defined to be the locus of a point P which satisfies the condition that the sum of the distances from P to given points A and B is constant. If we choose coordinates so that A is $(-d, 0)$ and B is $(d, 0)$, then this just means that ellipses are the level curves of the function

$$f((\xi_1, \xi_2)) = \sqrt{(\xi_1 + d)^2 + \xi_2^2} + \sqrt{(\xi_1 - d)^2 + \xi_2^2}$$

This can be checked by simplifying the equation $f(x) = c$ (see below).

The relation between the two types of definition above (direct and implicit) will be examined in the next chapter—it involves the use of the inverse function theorem and its variants. Suffice it to say that we obtain an implicit representation from a parametric one "by eliminating t", whereby we must take care not to lose part of the curve.

Another possibility for specifying curves which can often lead to considerable simplifications in dealing with concrete examples is that of using other coordinates systems.

Example: Consider the circle $\xi_1^2 + \xi_2^2 = 1$. With respect to polar coordinates (r, θ) where

$$\xi_1 = r\cos\theta, \quad \xi_2 = r\sin\theta$$

the circle has the implicit representation $r = 1$ and the parametric representation

$$r(t) = 1, \quad \theta(t) = t \quad (t \in [0, 2\pi[).$$

Abstractly, we can describe this as follows: let ϕ be a mapping from a subset U of $\mathbf{R}^2$ into $\mathbf{R}^2$. Then if c is a curve in U, we define the curve $\phi^*(c)$ to be the curve with the parametrisation $t \mapsto \phi \circ c(t)$. If c is the zero-set $\{x : f(x) = 0\}$, then $\phi^*(c)$ is the zero-set of the function $f \circ \phi^{-1}$.

Another popular method of specifying curves is by the use of so-called bipolar coordinates. Here two fixed points x_1 and x_2 are chosen (the poles) and the coordinates (r_1, r_2) are the respective distances from the poles i.e.

$$r_1 = |x - x_1|, r_2 = |x - x_2|.$$

(Note that these two numbers do not determine the point uniquely – its mirror image in the line through x_1 and x_2 has the same coordinates. Hence this method is only appropriate for describing curves which are symmetric with respect to reflection in this line. Also two numbers (r_1, r_2) are the coordinates of a point if and only if their sum is greater than or equal to the distance between x_1 and x_2).

For example, if we take the two foci of an ellipse as poles, then the bipolar equation of the latter is

$$r_1 + r_2 = 2a.$$

Similarly, the bipolar equation of a hyperbola, with its foci as poles, is

$$r_1 - r_2 = \pm 2a.$$

In the above language, we are considering the mapping

$$\phi : x \mapsto (|x - x_1|, |x - x_2|).$$

Then if c is the line $\xi_1 + \xi_2 = 2d$, the ellipse is the pre-image of this curve under ϕ i.e. the curve $\phi^{-1}(c)$.

Many curves are obtained as the images of simple curves (e.g. lines and circles) under suitable analytic or meromorphic functions. For example, straight lines can be described as follows: let z_0 be the (complex number which describes the) point of reflection of the origin in the line L. Then the vectors z and $(z - z_0)$ have the same length and so there is a complex number ω where $\omega \in T = \{\omega \in \mathbf{C} : |\omega| = 1\}$ with $z - z_0 = \omega z$. This simplifies to the equation

$$z = \frac{z_0}{1 - \omega}$$

and so the line is the image of T under the mapping

$$\omega \mapsto \frac{z_0}{1 - \omega}.$$

Sometimes it is more convenient to consider the **pre-image** $\phi^{-1}(c)$ of a curve c under a suitable mapping ϕ. For example, if c is defined implicitly by the equation $f = 0$, then its pre-image is the zero-set of the composed function $f \circ \phi$. The commonest examples are obtained by taking the preimages of the coordinate lines ($\Re\, z$ = constant,$\Im\, z$ = constant) or circles ($|z|$ = constant) under holomorphic mappings ϕ. Suitable candidates for ϕ are

$$z \mapsto z^n, z \mapsto \exp z, z \mapsto \frac{1}{2}\big(z + \frac{1}{z}\big).$$

For example, the preimages of the axes $\xi_1 = c$ resp. $\xi_2 = d$ under the mapping $z \mapsto z^2$ are the curves

$$\xi_1^2 - \xi_2^2 = c$$

resp.

$$2\xi_1\xi_2 = d.$$

(Note that these form two mutually orthogonal families of hyperbolas. In fact, families of curves generated in this way – i.e. as the preimages of the coordinate

axes under an analytic mapping – are always orthogonal. The reader is invited to ponder why this is the case).

Another (related) connection between complex numbers and curves is provided by the so-called **Schwarz function** of a curve. Consider firstly a curve in the plane given by the implicit equation $f(\xi_1, \xi_2) = 0$. If we identify the plane again with the set of complex numbers $\mathbf{C}$, then we can rewrite this equation in the form $\phi(z, \bar{z}) = 0$ for a suitable function ϕ of two complex variables (in fact,

$$\phi(z, \bar{z}) = f(\frac{z + \bar{z}}{2}, \frac{z - \bar{z}}{2i})).$$

Assuming that ϕ has reasonable properties (we will not concern ourselves here with the precise details which again involve the implicit function theorem), then we can solve the above equation to obtain one of the form $z = S(\bar{z})$ which expresses z explicitly as a function of $\bar{z}$. S is called the **Schwarz function** of the curve. For example, the straight line

$$a\xi_1 + b\xi_2 + c = 0$$

has the Schwarz function

$$S(z) = -\frac{(a - ib)z - 2c}{a + ib}$$

as the reader can verify. Similarly, the Schwarz function of the unit circle is $S(w) = \frac{1}{w}$.

Curves and differential equations: It is often helpful to use physical interpretations in visualising curves. For example, if we have a curve represented in parametric form $x = c(t)$ then we can regard t as a time variable and $c(t)$ as the position of the particle at time t so that the curve describes its motion in the plane. More generally, the coordinates of x can represent generalised coordinates in some phase space, for example, in the mathematical formulation of Newtonian mechanics, the vector x could represent in the first coordinate the position of a particle moving with one degree of freedom and in the second coordinate velocity. Such curves arise typically as solutions of differential equations of the form

$$\dot{x} = f(x, t)$$

where $x = (\xi_1(t), \xi_2(t))$. The above equation is thus equivalent to the system

$$\dot{\xi}_1 = f_1(\xi_1, \xi_2, t) \quad \dot{\xi}_2 = f_2(\xi_1, \xi_2, t).$$

Example: Consider a particle with one degree of freedom. If its position is represented in some coordinate system by the variable $\xi(t)$, then the general form of Newton's equation prescribes a second order ordinary differential equation of the form

$$\ddot{\xi}(t) = F(\xi(t), \dot{\xi}(t), t)$$

for a suitable function F. If we introduce the vector function

$$x(t) = (\xi_1(t), \xi_2(t))$$

where $\xi_1(t) = \xi(t)$ and $\xi_2(t) = \dot{\xi}(t)$, then this becomes

$$\dot{x}(t) = f(x(t), t)$$

where $f_1(x,t) = \xi_2$ and $f_2(x,t) = F(\xi_1, \xi_2, t)$. The solutions of these equations are then trajectories in the phase space of the particle.

We illustrate this with three simple examples:
I. Free fall: This corresponds to the equation $\ddot{\xi} = -g$. The corresponding system is

$$\dot{\xi}_1 = \xi_2 \quad \dot{\xi}_2 = -g.$$

II. Movement in a gravitational field emanating from a planet:

$$\ddot{\xi} = \frac{-gr_0}{(\xi + r_0)^2}$$

i.e.

$$\dot{\xi}_1 = \xi_2 \quad \dot{\xi}_2 = \frac{-gr_0}{(\xi_1 + r)^2}.$$

III. A weight under the action of a spring: The second order equation is $\ddot{\xi} = -\alpha^2 \xi$ with corresponding system

$$\dot{\xi}_1 = \xi_2, \quad \dot{\xi}_2 = -\alpha^2 \xi_1.$$

Note that all of these equations can be written in the form

$$\ddot{\xi} = F(\xi) = -\frac{\partial U}{\partial \xi}$$

where $U = -\int_{\xi_0}^{\xi} F(t)dt$. (In the above cases we have $U(\xi) = g\xi$, $\frac{-gr_0}{\xi + r_0}$, *resp.* $\frac{\alpha^2 \xi^2}{2}$). In this case the solutions of the corresponding system

$$\dot{\xi}_1 = \xi_2 \quad \dot{\xi}_2 = F(\xi_1)$$

have the implicit representation $E(x) = c$ where E is the energy function

$$\frac{\xi_2^2}{2} + U(\xi_1)$$

(traditionally written $E = T + U$ where T is the **kinetic energy** and U is **potential energy**).

This leads to the following mathematical formulation: Let G be an open subset of the plane. A **vector field** on G is a mapping f from G into the plane, whereby we tacitly assume some regularity condition on the field, usually at least continuous differentiability. The field then defines a family of curves, the solutions of the differential equation $\dot{x} = f(x)$. The typical behaviour of the solutions can be described as follows: through every point x_0 of G there passes exactly one solution of this equation and this determines a covering of G by a family of curves. If the vector field is the gradient of a function i.e. if f has the form $(\frac{\partial \phi}{\partial \xi_1}, \frac{\partial \phi}{\partial \xi_2})$ where ϕ is a scalar field, then the solution curves have the implicit representation $\phi(x) = c$.

Equations of the form $\dot{x} = f(x)$ i.e. where the right hand side does not depend explicitly on time are called **autonomous**. Then if x is a solution, so are the translated curves x_c where $x_c(t) = x(t - c)$.

Examples of autonomous systems:

$$\begin{array}{lllllll}
\dot{\xi}_1 & = & \xi_1, & \dot{\xi}_2 & = & -\xi_2 \\
\dot{\xi}_1 & = & -\xi_1, & \dot{\xi}_2 & = & -2\xi_2 \\
\dot{\xi}_1 & = & \xi_1, & \dot{\xi}_2 & = & \xi_1 + \xi_2 \\
\dot{\xi}_1 & = & \xi_1, & \dot{\xi}_2 & = & -\xi_2 \\
\dot{\xi}_1 & = & \xi_2, & \dot{\xi}_2 & = & -\xi_1 \\
\dot{\xi}_1 & = & -\xi_1, & \dot{\xi}_2 & = & -\xi_1 + \xi_2
\end{array}$$

Using the differential equation $\dot{x} = f(x)$ we can define a so-called **phase-flow** as follows: if $x \in \mathbf{R}^2$, we define $\phi_t(x)$ to be the value of the solution $x(t)$ of the equation at time t, starting from the initial value $x(0) = x$. Then we have the relation

$$\phi_{s+t}(x) = \phi_s(\phi_t(x)).$$

Typically the mappings ϕ_s are homeomorphisms of space i.e. we can regard the flow of the differential equation as generating a continuously changing deformation of space, the field lines being the trajectories of single points with respect to these deformations.

Example: For the equation $\ddot{\xi} = -\xi$ with corresponding system

$$\dot{\xi}_1 = \xi_2 \quad \dot{\xi}_2 = -\xi_1$$

we have the solution

$$x = (\xi_1 \cos t + \xi_2 \sin t, \xi_2 \cos t - \xi_1 \sin t)$$

with initial value $x(0) = (\xi_1, \xi_2)$. Here ϕ_t is the linear mapping with matrix

$$\begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}$$

(i.e. a rotation).

Often (as in the above example) these homeomorphisms ϕ_t are rigid i.e. isometries of the plane. This leads to the consideration of phase flows of the form

$$\phi_t = T_{u(t)} \circ f(t)$$

where f and u are smooth functions, the first taking its values in the family of isometries (i.e. orthogonal two-by-two matrices), the second in the plane. Then the trajectories take the form

$$c(t) = f(t)(x_0) + u(t)$$

for some starting point x_0. Important examples of such curves are the cycloid and its variants which we discuss below:

Remark: The differential equation $\dot{x} = f(x)$ is often written in classical notation in the form

$$Pdx + Qdy = 0$$

where

$$f_1(\xi_1, \xi_2) = \frac{1}{P(\xi_1, \xi_2)}, \quad f_2(\xi_1, \xi_2) = \frac{-1}{Q(\xi_1, \xi_2)}.$$

We shall simply regard the notation

$$Pdx + Qdy = 0$$

as a convenient short-hand for the system $\dot{x} = f(x)$ where f is as above.

Examples: Examples of such equations are

$$x(y^2 - 1)dx - y(x^2 - 1)dy = O$$

$$(ax + by)dx + (a_1x + b_1y)dy = 0$$

$$(1 + y^2)ydx + (1 + x^2)dy = 0$$

$$ydx - xdy = 0.$$

Systems of curves: In fact, curves seldom occur on their own but rather as members of suitable families. These can arise in various ways, of which the following examples are probably the most important:

a) a curve of the form $f(x) = 0$ is a member of the family $f(x) = c$ of contours of the landscape formed by the graph of f.

b) the solutions of the equation $Pdx + Qdy = 0$ typically form a one-parameter family of curves which cover some region of the plane.

c) orthogonal families: in applications, one often meets two families of curves which are mutually orthogonal. Such families can arise as follows: if the one family is the solution to the equation $Pdx + Qdy = 0$, then the second is the solution of $Qdx - Pdy = 0$.

d) If ϕ is a suitable map from the plane (or a suitable subset thereof) into the

plane, then ϕ maps families of curves into new ones. Thus we can obtain exotic families by applying suitable mappings to more humdrum systems (such as the lines parallel to the coordinate axes). Typical examples are obtained by using holomorphic functions (see above).

We conclude these informal remarks about curves with a list of some classical examples, grouped together according to the most convenient method of describing them.

A. **Curves as level surfaces:** As mentioned above, these often arise as the loci of points moving under some restraint which can be interpreted as a condition of the form $f(x) = c$ on the coordinates of the point.

I. The ellipse: This is often defined as the locus of a point which moves in such a way that the sum of its distances from two fixed points is constant (c.f. figure 4). For reasons which will soon become apparent, we now take the two fixed points (the foci of the ellipse) to be $(ae, 0)$ and $(-ae, 0)$, the constant to be $4a^2$. The equation then takes on the form

$$\sqrt{(\xi_1 + ae)^2 + \xi_2^2} + \sqrt{(\xi_1 - ae)^2 + \xi_2^2} = 4a^2$$

which simplifies to

$$\frac{\xi_1^2}{a^2} + \frac{\xi_2^2}{b^2} = 1$$

where $b^2 = a^2(1 - e^2)$.

II. The parabola (figure 4): This is the locus of a point whose distances from a given point F (the focus) and a given line L (the directrix) are equal. If we take F to be the point $(a, 0)$ and L to be the line $\xi_1 = -a$, we get the equation

$$(\xi_1 - a)^2 + \xi_2^2 = (\xi_1 + a)^2$$

which simplies to the familiar form $\xi_2^2 = 4a\xi_1$.

III. Cassini's ovals (figure 2: These were so named after being used by the astronomer Cassini in his investigation of the two body problem (earth-sun system). They are defined as the locus of a point P which moves in such a way that the product of its distances from two fixed points (the poles) is constant. If we take $(-a, 0)$ and $(a, 0)$ as the poles, we get the equation

$$((\xi_1 - a)^2 + \xi_2)^2((\xi_1 + a)^2 + \xi_2^2) = b^4$$

which simplifies to

$$(\xi_1^2 + \xi_2^2 + a^2)^2 = b^4 + 4a^2\xi_1^2.$$

These are closed, non-self-intersecting curves for $a < b$. For $a = b$ the curve is a figure of eight (the Leminiscate of Bernoulli) and for $a \geq b$ it splits up into two loops.

IV. The Lamé curves: This is the family of curves with implicit equations

$$\left(\frac{\xi_1^2}{a^2}\right)^n + \left(\frac{\xi_2^2}{b^2}\right)^n = 1.$$

(For each value of n between zero and infinity we get a separate curve.) Of course, the case $n = 1$ is the ellipse. The case where $n = \frac{1}{3}$ is an interesting curve called the **asteroid** which we shall discuss later. The case $n > 2$ gives elegant oval shapes which are popular in art and design.

B. Curves defined by parametrisations resp. by movements in the plane:

I. The cycloid (figure 4): This is the path traced by a point on the cicumference of a circle which is rolled along a line. From figure 3 we see that the vector OP is the sum of the vectors OM and MP i.e. it is $(t, 1) - D_{-t}(0, 1)$ or $(t - \sin\, t, 1 - \cos\, t)$.

More generally, if we trace the path of the point with original coordinates x (i.e. not necessarily the path of the origin as above), then we merely replace the vector MP by $D_{-t}(x - (0, 1))$. This leads to the equation

$$c(t) = (t + \xi_1 \cos t + (\xi_2 - 1) \sin t, 1 - \xi_1 \sin t + (\xi_2 - 1) \cos t).$$

II. Epicycloids resp. hypocycloids. These are obtained as above but by rolling a smaller circle (of radius r) around a larger one (with radius R). We suppose that $R = nr$, whereby n need not be an integer. If the smaller circle is rolled around the exterior of the larger one, we get an epicycloid, otherwise a hypocycloid. The method used above leads to the equations

$$c(t) = ((n + 1)r \cos\, t - r \cos\,(n + 1)t, (n + 1)r \sin\, t - r \sin\,(n + 1)t)$$

for the epicycloid and

$$c(t) = r((n - 1) \cos t + \cos(n - 1)t, (n - 1) \sin nt - \sin(n - 1)t)$$

for the hypocycloid.

Two special cases are of particular interest:

III. The nephroid (figure 4): This is the epicycloid for the case where $n = 2$. It has parametrisation

$$c(t) = r(3 \cos t - \cos 3t, 3 \sin t - \sin 3t).$$

IV. The cardioid (figure 5): This is the epicycloid with $n = 1$. It has parametrisation

$$c(t) = r(2 \cos t - \cos 2t, 2 \sin t - \sin 2t).$$

C. **Curves defined by differential equations:**

Recall the general equation $\dot{x} = f(x)$. The most interesting things happen around zeros of the field f (these correspond to states of equilibrium of physical systems). For convenience, we shall assume that this takes place at $x = 0$ i.e. the equation has the form $\dot{x} = f(x)$ where $f(0) = 0$. Now if we assume that f is smooth, we can consider its Taylor development. This begins with the linear term $(Df)_0(x)$.

Hence, neglecting the terms of higher order, we obtain, as an approximation to the original equation, one with a **linear** field f on the right hand side i.e. a system of the form $\dot{x} = Ax$ where A is a two by two matrix (the Jacobi matrix of f at zero). It is plausible (and with the usual suitable restrictions even true) that the solution to this linear equation will provide – at least in a neighbourhood of the origin – a good approximation to the solution of the general equation. For this reason, we shall confine our attention here to such linear systems. For such equations, one can give a complete description of the solutions. For reasons which will be explained shortly, we begin by considering the following four cases:
a)

$$A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

where λ_1 and λ_2 are distinct reals.
b)

$$A = \begin{bmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{bmatrix}.$$

c)

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

for λ real.
d)

$$A = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}$$

where α and β are real.
The corresponding solutions are
a) $\xi_1 = c_1 e^{\lambda_1 t},\ \xi_2 = c_2 e^{\lambda_2 t}$.
b) $\xi_1 = (c_1 + tc_2)e^{\lambda t},\ \xi_2 = c_1 e^{\lambda t}$.
c) as a) with equal lambdas.
d) $r(t) = r_0 e^{\alpha t},\ \theta(t) = \beta t + \theta_0$.
See figure 7.

In order to justify the choice of the above four types of matrices, we recall some elementary facts from linear algebra. Since our reasoning is completely general, we might just as well deal with the n-dimensional case i.e. with equations of the form $\dot{x} = Ax$ where A is an n by n matrix and x is a function with values in $\mathbf{R}^n$. Consider firstly the situation where A is diagonalisable i.e. there is an invertible matrix P so that $PAP^{-1} = D$ where

$$D = \text{diag}\,(\lambda_1, ..., \lambda_n).$$

Then the equation can be written in the form $P\dot{x} = DPx$ i.e. $\dot{y} = Dy$ where $y = Px$. Of course, this has solution

$$y(t) = (e^{\lambda_1 t}\eta_1, \ldots, e^{\lambda_n t}\eta_n)$$

with $y(0) = (\eta_1, ..., \eta_n)$.
The solution of the original equation is then $x = P^{-1}y$.

Similar techniques can be used when the matrix is not diagonalisable - then one must use the Jordan or the rational canonical form to reduce to simpler cases. For us it suffices to remark that a linear change of variables can always be found which reduces to the case where A has a suitable simple form. In the case of two by two matrices, only the four types considered above can occur. They correspond to the situations where A has two distinct real eigenvalues, two coincident real eigenvalues or a pair of complex-conjugate eigenvalues.

D. **Curves defined by holomorphic functions:**
Here we return to the topic of epi- and hypocycloids. If we refer to figure 6, we see that the complex number z corresponding to the vector OP has the form

$$z = OQ + QP = r(n-1)e^{i\theta} + re^{-i\theta(n-1)}.$$

Hence if we write w for the complex number $e^{i\theta}$, we see that the hypocycloid is the image of the unit circle T under the holomorphic mapping $z = \phi(w)$ where

$$\phi(w) = r((n-1)w + w^{1-n}).$$

It is interesting to note that our general condition on the mapping ϕ in order to be able to compose it with curves without introducing a singularity (i.e. that it be locally a diffeomorphism) fails when $\phi'(w) = 0$. In our case

$$\phi'(w) = \frac{r(n-1)(w^n - 1)}{w^n}$$

and in general there are cusps at the points where this expression vanishes. For example, in the case where $n = 2$ (where the transforming function is $z = r(w + \frac{1}{w})$) we have two cusps (in fact, this curve is a part of a straight line - a fact which is used in engineering). For $n = 3$ we have three cusps. The equation of this curve is

$$z = r(2w + \frac{1}{w^2}).$$

(The curve is called a **deltoid.**)

In a similar manner, one can calculate that the equation of the epicycloid is

$$z = r((n+1)w + w^{n+1}).$$

For $n = 1, r = 1$ we get $z = 2w - w^2$ which is the equation of the **cardioid.**

E. Curves defined by Schwarz functions:
I. The straight line through z_1, z_2 has Schwarz function:

$$S(z) = \frac{\bar{z}_1 - \bar{z}_2}{z_1 - z_2}z + \frac{z_1\bar{z}_2 - z_2\bar{z}_1}{z_1 - z_2}.$$

II. The circle with centre z_0 and radius r. The Schwarz function is

$$S(z) = \frac{r^2}{z - z_0} + \bar{z}_0.$$

III. The ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$:

$$S(z) = \frac{a^2 + b^2}{a^2 - b^2} z^2 + \frac{2ab}{b^2 - a^2} \sqrt{z^2 + b^2 - a^2}.$$

F. Curves in polar coordinates:
I. The Rhodoneae: This is the family of curves with polar equations

$$r = a \cos k\theta resp. \quad r = a \sin k\theta$$

where k is a parameter (not necessarily an integer). For integral values, they are rose-like curves. For example, the case $k = 4$ is a four petalled flower-shape—known as the **quadrifolium.**
II. The spirals: "Spiralïs a generic name for curves of the form $r = f(\theta)$ where f is usually (but not always) positive and monotone. The best known examples are
1) the spiral of Archimedes with equation $r = a\theta$;
2) the spiral of Fermat with equation $r^2 = a^2\theta$;
3) the spiral or Lituus with equation $r^2 = \frac{a^2}{\theta}$;
4) the hyperbolic spiral with equation $r = \frac{a}{\theta}$;
5) the equiangular spiral with equation $r = a^\theta$;
6) the sinuisoidal with equation $r^n = a^n \cos n\theta$ (n a parameter).
(See figue 8).

1.1 Surfaces

We now discuss briefly surfaces, more precisely, two dimensional surfaces in three-space. Once again, there are several ways of describing them and we shall concentrate on the following two:
The implicit definition: surfaces are the zero-sets of smooth functions defined on suitable subsets of $\mathbf{R}^3$. For example, the sphere is the zero-set of the function

$$f(x) = \xi_1^2 + \xi_2^2 + \xi_3^2 - 1.$$

As in the case of curves, such surfaces can be regarded as the level surfaces of a scalar field in space. For example if the scalar field represents a potential, then these are the **equipotentials.** A consideration of concrete examples will speedily persuade the reader that singularities occur only when the gradient of the scalar function f vanishes. (Think of the vertex of the cone $\xi_3^2 = \xi_1^2 + \xi_2^2$.)

The parametric definition: Surfaces are the images of (open subsets of) $\mathbf{R}^2$ under smooth mappings from the plane into space. For example the mapping

$$\phi : (u, v) \mapsto (u, v, \sqrt{1 - u^2 - v^2})$$

which is defined for $u^2 + v^2 < 1$ describes a hemisphere.

Once again, simple examples suggest that singularities can only occur when the rank of the derivative $D\phi$ of ϕ is not maximal i.e. is either 0 or 1. Geometrically, this means that the two vectors $D_1\phi$ and $D_2\phi$ (the partial derivatives of ϕ) are proportional.

Examples of surfaces:

I. Landscapes: These are surfaces of the form $\xi_3 = f(\xi_1, \xi_2)$ i.e. the graph of a smooth function defined on the plane. Such surfaces have the implicit representation $F = 0$ where F is the function

$$x \mapsto \xi_3 - f(\xi_1, \xi_3)$$

and the parametric representation

$$\phi(u, v) = (u, v, f(u, v)).$$

II. Surfaces of revolution (figure 12): these are surfaces obtained by rotating a curve

$$c(u) = (0, h(u), k(u))$$

in the (y, z)-plane around the z-axis. They are parametrised by the function

$$\phi : (u, v) \mapsto (h(u)\cos v, h(u)\sin v, k(u)).$$

If the curve has the implicit form $F(\xi_2, \xi_3) = 0$, then the implicit equation of the surface is

$$F((\xi_1^2 + \xi_2^2)^{\frac{1}{2}}, \xi_3) = 0.$$

A simple and important example of a surface of revolution is the standard cone which is obtained by rotating the diagonal in the (y, z)-plane about the z-axis. It has thus the implicit representation $\xi_3^2 = \xi_1^2 + \xi_2^2$ and the parametrisation $\phi(u, v) = (u\cos v, u\sin v, u)$.

III. The sphere (figure 9): this has implicit equation

$$\xi_1^2 + \xi_2^2 + \xi_3^2 = 1.$$

and parametrisation

$$\phi(u, v) = (\cos u \cos v, \cos u \sin v, \sin u)$$

as surface of revolution generated by the unit circle in the (y, z)-plane. (parametrisations of the sphere are of particular interest since they form the basis of cartography).

IV. The torus: this is also a surface of revolution, this time of a circle as in the diagram.

The torus has parametrisation

$$\phi(u, v) = ((a + b\cos u)\cos v, (a + b\cos u)\sin v, b\sin u)$$

where $a > b > 0$, a and b being the radii of the circles indicated in the diagram. The implicit equation is

$$\left(\sqrt{\xi_1^2 + \xi_2^2} - a\right)^2 + \xi_3^2 = b^2.$$

V. The cone (figure 10): this is again a surface of revolution, with parametrisation

$$\phi(u, v) = (u \cos v, u \sin v, u)$$

and implicit equation $\xi_3^2 = \xi_1^2 + \xi_2^2$.

Note that at the points $u = \pm\frac{\pi}{2}$ resp. $u = v = 0$, the parametrisations of the sphere and the cone are not regular. In the second case, but not in the first, this corresponds to a real singularity of the underlying figure.

VI. The helicoid (figure 11), with parametrisation

$$\phi(u, v) = (u \cos v, u \sin v, v) \quad (u, v \in \mathbf{R}).$$

VII. The Möbius strip has parametrisation

$$\phi(u, v) = (\cos u + \sin u \cos u, \sin u + v \sin u \sin v, v \cos u).$$

VII. Cylinders: these have equations $f(x) = 0$ where the function has the form $f(\xi_1, \xi_2, \xi_3) = g(\xi_1, \xi_2)$ for a function g of two variables (i.e. f is independent of the third variable). This surface is **the cylinder** over the plane curve formed by the zero-set of g.

If ϕ is the parametrisation of a surface, then the curves

$$u \to \phi(u, v_0) \quad v \to \phi(u_0, v)$$

obtained by holding v resp. u fixed are called the (curvilinear) coordinate lines on the surface (they are just the images of the Cartesian coordinate lines under the parametrisations). For example, for the parametrisation of the sphere given above, they are the familiar lines of latitude and longitude. At a given point on the surface, the two tangents to these lines span the **tangential plane** there. More precisely, if ϕ_1 and ϕ_2 denote the partial derivatives of ϕ and we introduce the mapping

$$\mathbf{N}(u, v) = \frac{\phi_1(u, v) \times \phi_2(u, v)}{|\phi_1(u, v) \times \phi_2(u, v)|}$$

(which is called the **Gaußian mapping**), then $\mathbf{N}$ is the unit normal to the surface at $\phi(u, v)$ and the tangential plane there has he equation

$$(x - \phi(u, v)|\mathbf{N}(u, v)) = 0.$$

The behaviour of $\mathbf{N}$ and its derivative provide information on the geometry of the surface in a neighbourhood of the given point. This topic will be discussed in more detail later.

2 CURVES IN THE PLANE

In this chapter we shall bring a more systematic and general theory of curves. The emphasis will be on structural properties rather than on the special character of particular curves. We begin with the formal definition of parametrised curves in the plane:

Definition: A parametrised C^r-curve in $\mathbf{R}^2$ is an r-times continuously differentiable function c from an interval I in $\mathbf{R}$ into $\mathbf{R}^2$. The parametrisation is **regular** if $\dot{c}(t) \neq 0$ for all t.

Of course, such a parametrisation contains more information than one usually associates with a geometric curve (if one thinks of a curve as the path of a moving particle, then the parametrisation also tells us the speed of the particle at a given moment). Hence we introduce the following concept of equivalence between parametrisations. Two parametrisations c_1 and c_2 are **equivalent** if there is a C^r-bijection $\phi : I_1 \to I_2$ (where the I's are the respective domains of definition) such that $\dot{\phi}(t) > 0$ for each $t \in I_1$ and $c_2 = c_1 \circ \phi^{-1}$. Such a function ϕ is called a **reparametrisation.** Note that the condition on ϕ ensures that its inverse is also C^r. The positivity condition on the derivative means that we are prohibiting reversals of direction.

Of course it would be horribly tedious to distinguish between parametrisations, equivalence classes thereof and the various degrees of regularity. Hence we shall often simply employ the word "curveänd leave it to the common sense of the reader to deduce from the context in which precise sense it is being used. If there is any danger of confusion we shall be more precise in our use of terminology.

The simplest version of the implicit function theorem implies the following: if $c : I \to \mathbf{R}^2$ is a regular C^r-curve (for $r \geq 1$) and t_0 is in the interior of I, then there is a neighbourhood V of $c(t_0)$ in $\mathbf{R}^2$ and a diffeomorphism ψ from V onto a neighbourhood U of zero in $\mathbf{R}^2$ so that $\psi \circ c$ is equal to the curve $t \mapsto (t, 0)$. Note that this implies that locally the curve c is the zero set of a smooth function (namely the second component ψ_2 of ψ).

Particularly simple are curves of the form $c(t) = (t, f(t))$ i.e. the graphs of functions. In fact all regular curves are locally of this form (again a consequence of the implicit function theorem). More precisely, choose $t_0 \in I$. Since $\dot{c}(t_0) \neq 0$, either $\dot{c}_1(t) \neq 0$ or $\dot{c}_2(t) \neq 0$. Suppose that the former holds. Then there is a neighbourhood U of t_0 in I so that c_1 is invertible on U. Let ϕ be an inverse to the restriction of c_1 to U. Then under the reparametrisation ϕ, the curve has the form $u \to (u, c_2(\phi(u))$ on U.

Arc-length: If $c : [a, b] \to \mathbf{R}^2$ is a curve, its **length** is defined to be

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

Geometrically, the length is defined as follows: we choose a partition $(t_0, t_1, \ldots, t_n)$ of $[a, b]$ and consider the broken line constructed by joining successively the points $(c(t_0), c(t_1), \ldots, c(t_n))$. The length of the curve is the limit of the lengths of these lines as the partition becomes finer. Since we shall not require this result, we will not bother to prove it. This can be done easily by writing out explicitly the length of the chain and noting that an application of the mean value theorem displays it as a Riemann sum for the above integral.

We can use the notion of arc length to introduce a natural parametrisation for curves. If we regard a curve as the path of a moving particle, then there is one form of motion which obviously enjoys a privileged position among all equivalent ones – that for which the speed is uniform. This means that the particle arrives at a point in a time which is proportional to its distance (along the curve) from its starting point. Hence we use the reparametrisation ϕ where

$$\phi(t) = \int_a^t |\dot{c}(u)|\, du.$$

It is customary to write s for $\phi(t)$. The new parametrisation $c \circ \phi^{-1}$ is called the **parametrisation by arc-length** and is traditionally denoted by γ. Thus we have the relationship

$$\gamma(s) = c(t) \qquad (s = \phi(t)).$$

When we use the letter γ to denote the parametrisation of a curve in future, then we are tacitly assuming that it is parametrised by arc-length. Another generally employed convention is to use dashes to denote the derivative with respect to s and Newtonian dots for differentiation with respect to t (thus $\gamma'(s)$ but $\dot{c}(t)$).

If we differentiate the expression $\gamma(s) = c(t)$ and use the chain rule (recalling that $\dot{\phi}(t) = |\dot{c}(t)|$), then we see that

$$\gamma'(s) = \frac{\dot{c}(t)}{|\dot{c}(t)|}$$

(i.e. the derivative of γ is the normalised version of the derivative of c).

Now the difference quotients

$$\frac{\gamma(s+h) - \gamma(s)}{h}$$

which define the derivative of γ at the point s represent the chord from γs to $\gamma(s + h)$, almost normalised (since the length of the chord is approximately h when h is small). Hence the unit vector $\gamma'(s)$ is the **unit tangent vector** to the curve at s. We denote it by $\mathbf{T}_\gamma(s)$ or, if there is no danger of confusion, simply by $\mathbf{T}(s)$. In terms of a general parametrisation c (i.e. not necessarily parametrisation by arc-length), we define

$$\mathbf{T}_c(t) = \frac{\dot{c}(t)}{|\dot{c}(t)|}.$$

We now wish to define the curvature of a curve at a point s. Intuitively, it is related to the rate of change of the tangent vector $\mathbf{T}$. We define it to be the reciprocal $\frac{1}{R}$ of the radius of that circle which approximates the curve best at the given point. More precisely, we shall show that under suitable circumstances, if s_1, s_2, s_3 are near s, then $\gamma(s_1), \gamma(s_2), \gamma(s_3)$ are not collinear and so there is a circle with centre $C(s_1, s_2, s_3)$ through them. Furthermore, we shall show that as the three points tend to s, then the centres $C(s_1, s_2, s_3)$ tend to a point. We then define the circle through $\gamma(s)$, with centre at this limit, to be the **osculating circle** of the curve at the given point. Its centre (resp. radius) are called the **centre of curvature** resp. **radius of curvature** at s. The inverse of the radius is called the **curvature** at s. The mathematics behind these definitions is contained in the following Proposition:

Proposition 2.1 *Let γ be a C^2-curve and suppose that a point s is such that $\gamma''(s) \neq 0$. Then there is a neighbourhood U of s so that $\gamma(s_1), \gamma(s_2), \gamma(s_3)$ are not collinear if s_1, s_2, s_3 are distinct points in U. As s_1, s_2, s_3 tend to s, the circle through $\gamma(s_1), \gamma(s_2), \gamma(s_3)$ converges to a circle through $\gamma(s)$ with radius $|\gamma''(s)|^{-1}$ which is tangential to the curve at $\gamma(s)$.*

Proof. Denote the centre of this circle by $C(s_1, s_2, s_3)$ as in the text above and consider the function

$$f : s \mapsto |\gamma(s) - C(s_1, s_2, s_3)|^2 = (\gamma(s) - C(s_1, s_2, s_3)|\gamma(s) - C(s_1, s_2, s_3)).$$

Then

$$f'(s) = 2(\gamma'(s)|\gamma(s) - C(s_1, s_2, s_3))$$

and

$$f''(s) = 2\big(\gamma''(s)|\gamma(s) - C(s_1, s_2, s_3)\big) + 2.$$

By the mean value theorem, there are points ξ_1, ξ_2, ξ_3 near s so that $f'(\xi_1) = f'(\xi_2) = 0$ and $f''(\xi_3) = 0$ i.e.

$$(\gamma'(\xi_i)|\gamma(\xi_i) - C(s_1, s_2, s_3)) = 0$$

for $i = 1, 2$ and

$$(\gamma''(\xi_3)|\gamma(\xi_3) - C(s_1, s_2, s_3)) = -1$$

(we are assuming, for convenience, that $s_1 < s_2 < s_3$).

If $C(s_1, s_2, s_3)$ has a limit C when the three points tend to s, we get $(\gamma'(s)|\gamma(s) - C) = 0$ (or $(\mathbf{T}(s)|\gamma(s) - C) = 0$). This means that the limiting circle is tangential to the curve. Furthermore we have

$$(\gamma''(s)|\gamma(s) - C) = -1.$$

Now if $\gamma''(s)$ is not a multiple of $\gamma'(s)$, then these equations determine C. In fact, $\gamma'(s)$ is perpendicular to $\gamma''(s)$ and both are non-zero by hypothesis. Hence the equations

$$(\gamma'(s)|\gamma(s) - C) = 0$$

and

$$(\gamma''(s)|\gamma(s) - C) = -1$$

determine C uniquely. In fact, $\gamma(s) - C = a\gamma''(s)$ where $a = \frac{-1}{|\gamma''(s)|^2}$ and

$$|\gamma(s) - C| = |a||\gamma''(s)| = \frac{1}{|\gamma''(s)|}.$$

We now prove the statement concerning the non-collinearity of $\gamma(s_1), \gamma(s_2)$ and $\gamma(s_3)$. The rest of the result follows then form the calculations above. Suppose there are distinct points s_1, s_2 and s_3 which are arbitrarily close to s with $\gamma(s_1), \gamma(s_2)$ and $\gamma(s_3)$ collinear. Then by the mean value theorem, there are points ξ_4, ξ_5 with ξ_4 between s_1 and s_2 and ξ_5 between s_2 and s_3, so that $\mathbf{T}(\xi_4)$ and $\mathbf{T}(\xi_5)$ are parallel to this line and hence to each other. Then $\mathbf{T}(\xi_4) = \pm\mathbf{T}(\xi_5)$. By continuity, we must have $\mathbf{T}(\xi_4) = \mathbf{T}(\xi_5)$ if we are near enough to s. This implies the existence of a ξ_6 between ξ_4 and ξ_5 with $\mathbf{T}'(\xi_6) = \gamma''(\xi_6)$ parallel to $\mathbf{T}(\xi_4) = \gamma'(\xi_4)$. In the limit, this implies that $\gamma''(s)$ is parallel to $\gamma'(s)$. But $\gamma''(s) \perp \gamma'(s)$ and $\gamma''(s) \neq 0$ which is a contradition.

In order to show that the centres $C(s_1, s_2, s_3)$ converge, we note that the equations

$$(\gamma'(\xi_1)|\gamma(\xi_1) - C(s_1, s_2, s_3)) = 0$$

and

$$(\gamma''(\xi_3)|\gamma(\xi_3) - C(s_1, s_2, s_3)) = -1$$

can be written in the matrix form

$$A(\xi_1, \xi_3)C(s_1, s_2, s_3) = \begin{bmatrix} (\gamma'(\xi_1)|\gamma(\xi_1)) \\ 1 + (\gamma''(\xi_3)|\gamma(\xi_3)) \end{bmatrix}$$

where

$$A(s, t) = \begin{bmatrix} \gamma_1'(s) & \gamma_2'(s) \\ \gamma_1''(t) & \gamma_2''(t) \end{bmatrix}$$

If we let the point converge to s, then we see that since $A(\xi_1, \xi_3)$ converges to

$$\begin{bmatrix} \gamma_1'(s) & \gamma_2'(s) \\ \gamma_1''(s) & \gamma_2''(s) \end{bmatrix}$$

and the right hand side converges to

$$\begin{bmatrix} (\gamma'(s)|\gamma''(s)) \\ 1 + (\gamma''(s)|\gamma(s)) \end{bmatrix}$$

then $C(s_1, s_2, s_3)$ converges to

$$A(s,s)^{-1} \begin{bmatrix} (\gamma'(s)|\gamma(s)) \\ 1 + (\gamma''(s)|\gamma(s)) \end{bmatrix}.$$

■

The **unit normal** $\mathbf{N}_\gamma(s)$ (or simply $\mathbf{N}(s)$) to the curve at s is the image $D_{\frac{\pi}{2}}\mathbf{T}_\gamma(s)$ of the tangent vector under the operator of rotation through 90 degrees (i.e. the mapping $(\xi_1, \xi_2) \to (-\xi_2, \xi_1)$). Thus $(\mathbf{T}(s), \mathbf{N}(s))$ forms a right handed orthogonal system for the plane.

From the above calculations we know that $\gamma''(s)$ is perpendicular to $\mathbf{T}(s)$ and so is some multiple of $\mathbf{N}(s)$. We can thus define the curvature $\kappa(s)$ of the curve at s by means of the equation

$$\gamma''(s) = \kappa(s)\mathbf{N}(s).$$

In other words, the absolute value of the curvature is just the length of the vector $\gamma''(s)$. This means that the curvature is, up to sign, the reciprocal of the radius of curvature. The sign of κ has the following geometrical significance:

$\kappa > 0$ means that the curve is curving towards $\mathbf{N}$;
$\kappa < 0$ means that the curve is curving away from $\mathbf{N}$.
(in both cases in the direction of increasing s.)

We have the following formula for the curvature function:

$$\kappa(s) = \det \begin{bmatrix} \gamma_1'(s) & \gamma_1''(s) \\ \gamma_2'(s) & \gamma_2''(s) \end{bmatrix}.$$

(The above determinant is precisely the signed area of the rectangle spanned by the vectors $\gamma'(s)$ and $\gamma''(s)$. For this latter expression is

$$-(\gamma'(s)|D_{\frac{\pi}{2}}\gamma''(s)) = -(\mathbf{T}_\gamma(s)| - \kappa(s)\mathbf{T}_\gamma(s)) = \kappa(s).)$$

The above information can be conveniently expressed in the equations:

$$\mathbf{T}'(s) = \kappa(s)\mathbf{N}(s) \qquad \mathbf{N}'(s) = -\kappa(s)\mathbf{T}(s)$$

which are known as the **Frenet formulae**.
(In matrix form

$$\begin{bmatrix} \mathbf{T}_1' & \mathbf{T}_2' \\ \mathbf{N}_1' & \mathbf{N}_2' \end{bmatrix} = \begin{bmatrix} 0 & \kappa \\ -\kappa & 0 \end{bmatrix} \begin{bmatrix} \mathbf{T}_1 & \mathbf{T}_2 \\ \mathbf{N}_1 & \mathbf{N}_2 \end{bmatrix}.$$

PROOF. The first equation is the definition of the curvature function. For the second one, we simply differentiate the equation $\mathbf{N} = D_{\frac{\pi}{2}}\mathbf{T}$ to get

$$\mathbf{N}' = D_{\frac{\pi}{2}}\mathbf{T}' = D_{\frac{\pi}{2}}\kappa\mathbf{N} = \kappa D_{\frac{\pi}{2}}D_{\frac{\pi}{2}}\mathbf{T} = -\kappa\mathbf{T}.$$

■

In order to be able to calculate with general curves (which are not usually presented in the convenient form of being parametrised by arc length), we construe these formulae in terms of the parametrisation $c = \gamma \circ \phi$. Firstly, we define the curvature and normal to c in the natural way i.e.

$$\kappa_{c(t)} = \kappa_\gamma(s), \quad \mathbf{N}_c(t) = \mathbf{N}_\gamma(s).$$

Then, by the chain rule,

$$\dot{c}(t) = \gamma'(s)\dot{\phi}(t)$$
$$\ddot{c}(t) = \gamma''(s)\dot{\phi}(t)^2 + \gamma'(s)\ddot{\phi}(t).$$

Using the fact that the curvature is the determinant of the matrix with the vectors $\gamma'(s)$ and $\gamma''(s)$ as columns, we obtain the formula

$$\kappa_c(t) = \frac{\dot{c}_1(t)\ddot{c}_2(t) - \ddot{c}_1(t)\dot{c}_2(t)}{(\dot{c}_1(t)^2 + \dot{c}_2(t)^2)^{\frac{3}{2}}}.$$

Also

$$\dot{\mathbf{T}}_c(t) = \mathbf{T}_{\gamma'}(\phi(t))\dot{\phi} = \mathbf{T}_{\gamma'}(s)|\dot{c}(t)| = \kappa_\gamma(s)\mathbf{N}_\gamma(s)|\dot{c}(t)|$$
$$= \kappa_{c(t)}\mathbf{N}_{c(t)}|\dot{c}(t)|$$

and so the Frenet formulae take on the form:

$$\dot{\mathbf{T}} = |\dot{c}|\kappa\mathbf{N} \qquad \dot{\mathbf{N}} = -|\dot{c}|\kappa\mathbf{T}.$$

A rather simple calculation shows that, as one would expect, the only curves with identically zero curvature (resp. with constant, but non-zero curvature) are straight lines (resp. circles). However, rather than actually carry this out, we shall prove a more general result, namely that the curvature determines the curve up to its position in the plane (i.e. two curves with the same curvature functions are congruent). On the other hand, as we shall show, any continuous function can be the curvature function of a curve.

Proposition 2.2 *Let f be a continuous, real-valued function on the interval $[0, L]$. Then there is a curve γ defined on this interval which has f as its curvature function. If $\tilde{\gamma}$ is a second curve which also has f as its curvature function, then the curves γ and $\tilde{\gamma}$ are congruent.*

Proof. Consider the differential equations

$$\mathbf{T}_1' = -f(s)\mathbf{T}_2, \quad \mathbf{T}_2'(s) = f(s)\mathbf{T}_1.$$

By a standard existence theorem for linear differential equations, there is a solution $\mathbf{T}$ of the above equations which satisfies any suitable initial conditions. We

choose *any* such conditions, whereby $|\mathbf{T}(0)| = 1$. Then the length of the vector is always one since

$$(\mathbf{T}_1^2 + \mathbf{T}_2^2)' = 2(\mathbf{T}_1\mathbf{T}_1' + \mathbf{T}_2\mathbf{T}_2') = -f(s)\mathbf{T}_1\mathbf{T}_2 + f(s)\mathbf{T}_1\mathbf{T}_2 = 0$$

and so the length of $\mathbf{T}$ is constant.

If we now take γ to be a primitive of $\mathbf{T}$, then by its very definition this is a curve, parametrised by arc-length, which satisfies the Frenet formulae, with κ replaced by f. Hence f is the curvature function of γ.

Now if $\tilde{\gamma}$ is a second such curve, then $\tilde{\mathbf{T}} = \tilde{\gamma}'$ is also a solution of the above differential equation (but with different inital values). If we choose a rotation D_θ about zero which maps $\mathbf{T}(0)$ onto $\tilde{\mathbf{T}}(0)$, then $D_\theta \circ \mathbf{T}$ is also a solution, this time with the same initial values as $\tilde{\mathbf{T}}$. Hence, by uniqueness, $\tilde{\mathbf{T}}$ coincides with $D_\theta \circ \mathbf{T}$. This implies the final statement of the theorem.

■

A disadvantage of the above definition of curvature is that it only applies to curves γ for which γ'' never vanishes. In particular, the curvature of a straight line is not defined. It should, of course, be zero. This can be rectified as follows. We say that a curve γ has a **moving frame** if there are continuous functions $\mathbf{T}, \mathbf{N}$ from $[0, L]$ into the plane and a continuous real-valued function κ on $[0, L]$ so that the pair $(\mathbf{T}(s), \mathbf{N}(s))$ forms a right-handed orthonormal basis for each s and the Frenet formulae

$$\mathbf{T}'(s) = \kappa(s)\mathbf{N}(s), \quad \mathbf{N}'(s) = -\kappa(s)\mathbf{T}(s)$$

hold. Also $\gamma'(s) = \mathbf{T}(s)$. κ is then called the curvature function of the curve. Under this definition, straight lines *do* have zero curvature.

It is clear that the angle between the tangent of a curve and the x-axis varies continuously along the curve. In order to state this precisely, we require the following Lemma:

Lemma 2.3 *Let f be a continuous function from an interval into the circle S^1. Then there is a continuous function $\tilde{f}$ from the interval into the real line so that $p \circ \tilde{f} = f$ where p is the mapping $t \mapsto (\cos t, \sin t)$ from $\mathbf{R}$ onto S^1. If $\tilde{g}$ is a second function with this property, then the difference $\tilde{f} - \tilde{g}$ is a constant function of the form $2\pi k$ for some $k \in \mathbf{Z}$.*

Proof. For convenience, we suppose that the interval of definition is the unit interval $[0, 1]$. By continuity, we can choose an $n \in \mathbf{N}$ so large that the range of f on each interval of the form $[\frac{k}{n}, \frac{k+1}{n}]$ lies in a half-circle of S^1. Choose $\tilde{f}(0)$ to be some point in $\mathbf{R}$ with $p(\tilde{f}(0)) = f(0)$. Now it is clear from the diagram how we should (even must) define $\tilde{f}$ on $[0, \frac{1}{n}]$. We then repeat this process to define $\tilde{f}$ successively on the intervals $[\frac{k}{n}, \frac{k+1}{n}]$.

If $\tilde{g}$ is a second such function, then the difference $\tilde{f} - \tilde{g}$ is a continuous function which takes its values in the set $\{2\pi k : k \in \mathbf{Z}\}$. But any interval in $\mathbf{R}$ is connected and hence so is its image in the latter set. The only connected subsets of discrete spaces are one-point sets and this implies that the difference is constant.

■

The following generalisation of this result can be proved similarly:

Proposition 2.4 *Let G be a subset of the plane which is starlike with respect to a point x_0. Then if f is a continuous mapping from G into S^1, there is a continuous lifting $\tilde{f}$ of f i.e. a function from G into $\mathbf{R}$ so that $p \circ \tilde{f} = f$.*

If we apply the above result to the tangent function $\mathbf{T}$ of a curve γ, then it guarantees the existence of a continuous function θ so that

$$\mathbf{T}(s) = (\cos\theta(s), \sin\theta(s))$$

(i.e. θ describes in a continuous way the angle between the tangent and the x-axis).

It follows from the above equation that

$$\gamma''(s) = (-\theta'(s)\sin\theta(s), \theta'(s)\cos\theta(s))$$

and so

Definition 2.5 *A* **first order differential form** *(or simply a* **one form***) on an open subset U of $\mathbf{R}^2$ is a mapping*

$$\omega : U \times \mathbf{R}^2 \to \mathbf{R}$$

which is linear in the second variable and smooth in the first one. In other words, it assigns to each point x of U an element of the dual of $\mathbf{R}^2$. The standard example is the differential df of a smooth function $f : U \to \mathbf{R}$ where

$$df(x) : y = (\eta_1, \eta_2) \mapsto D_1 f(x)\eta_1 + D_2 f(x)\eta_2$$

(i.e. $df(x)$ is the vector $(D_1 f, D_2 f)_x$ regarded as an element of the dual of $\mathbf{R}^2$ in the usual way).

In order to develop a more suggestive notation for differential forms, we write ξ_1 resp. ξ_2 for the functions

$$x \mapsto \xi_1 \quad x \mapsto \xi_2$$

on $\mathbf{R}^2$. Then the differentials $d\xi_1$ and $d\xi_2$ are constant. In fact, the pair $(d\xi_1, d\xi_2)$ is just the canonical basis for the plane and the above formula takes on the familiar form

$$df = D_1 f d\xi_1 + D_2 f d\xi_2$$

($df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy$ in classical notation). If the field f is interpreted as a potential function, then the above one-form corresponds to the field induced by the potential.

The general one form has then a representation

$$\omega(x) = a_1(x)d\xi_1 + a_2(x)d\xi_2$$

where a_1 and a_2 are smooth functions on the domain (in other words, they are just the coordinates of the form with respect to the canonical basis at a given point x).

An important example of a 1-form is given by the following formula:

$$\omega = \frac{-\xi_2}{\xi_1^2 + \xi_2^2}d\xi_1 + \frac{\xi_1}{\xi_1^2 + \xi_2^2}d\xi_2.$$

This is a one form on the punctured plane. We note for future reference that if U is any region in the punctured plane which is such that there is a ray h from 0 which misses U then we can define a smooth function θ on U so that $\theta(x)$ is the angle between the ray and the vector x (θ is any suitable branch of the complex function arg where we identify $x = (\xi_1, \xi_2)$ with the complex number $z = \xi_1 + i\xi_2$). On such a region, the above form ω is the differential of θ as the reader can verify.

We now define the curvilinear integral $\int_c \omega$ of a one form in a domain U where c is a curve with trace in U. This is done by means of the formula

$$\int_c \omega = \int_a^b \omega(c(t))\dot{c}(t)dt = \int_a^b a_1(c(t))\dot{c}_1(t) + a_2(c(t))\dot{c}_2(t)dt.$$

Note that this is invariant under a reprametrisation.

If ω is the differential df of a smooth function f then

$$\int_c \omega = f(c(b)) - f(c(a)).$$

For

$$\begin{aligned}\int_c df &= \int_a^b D_1 f(c(t))\dot{c}_1(t)dt + D_2 f(c(t))\dot{c}_2(t)dt \\ &= \int_a^b (f \circ c)'(t)dt = f(c(b) - f(c(a)).\end{aligned}$$

In particular, this implies that the integral is zero over a closed curve. This means that the field defined by a potential function is conservative.

On the other hand, if we calculate the integral of the particular ω prescribed above around the unit circle, then we obtain the value 2π. This shows that our form cannot be the differential of a smooth function on the punctured plane (in constrast to the situation described for regions which are missed by suitable rays).

We now consider a closed curve c in the punctured plane. We know from the above that there is a continuous function θ so that

$$\frac{c(t)}{|c(t)|} = (\cos\theta(t), \sin\theta(t)).$$

(this function θ is related to, but not identical with, the function θ mentioned above).

In fact, we claim that such a θ is $\theta(a) + \int_{c_t} \omega$ where $\theta(a)$ is a real number so that

$$\frac{c(a)}{|c(a)|} = (\cos\theta(a), \sin\theta(a)),$$

ω is the above form and c_t is the curve $c|_{[a,t]}$. In order to prove this we consider a partition $a = t_0 < t_1 < \ldots t_n = b$ of our interval which is such that the trace of the curve on each interval $[t_i, t_{i+1}]$ lies in a segment with opening $< \pi$ and centre 0. Then in this region ω has the form df (where f is a suitable branch of the argument funtion) and so the integral of the form along the segment is just the angle between $c(t_i)$ and $c(t_{i+1})$. Hence the integral of ω along c is the sum of those angles which sums to the angle between $c(a)$ and $c(b)$.

With this in mind, we define the **winding number** $w(c; 0)$ of a closed curve in the punctured plane with repect to 0 by means of the formula

$$w(c; 0) = \frac{1}{2\pi} \int_c \omega$$

where ω is the above form. Of course the above expression is a whole number.

More generally, we can define $w(c; a)$ where a is a point which does not lie on the trace of a curve c simply to be $w(c - a; 0)$.

We note some simple properties of the winding number:

a) it is independent of the parametrisation (since so is the curvilinear integral);

b) if a and b are points which do not lie on c and which can be joined by a continuous curve which does not cross c, then $w(c; a) = w(c; b)$ (for the winding number varies continuously along the curve and so is constant since it is integral-valued);

c) if c_1 and c_2 are curves, a a point not on either of them and the two curves are **homotopic** in the punctured plane, then $w(c_1; 0) = w(c_2; 0)$. The argument is as in c).

Now let c be a closed, simple curve which is such that $c'(a) = c'(b)$. Then the tangent mapping $\mathbf{T}$ can be regarded as a closed curve in S^1 and we define $r(c)$ – the **rotational index** of c – to be the winding number of $\mathbf{T}$ around 0.

Notice that if two curves c_1 and c_2 are **isotopic** i.e. if there is a smooth mapping $H : [a, b] \times [0, 1] \to \mathbf{R}^2$ so that $H(t, 0) = c_1(t)$ and $H(t, 1) = c_2(t)$ for $t \in [a, b]$ and $D_1(t, u)$ never vanishes, then the tangent curves $\mathbf{T}_1$ and $\mathbf{T}_2$ are homotopic under the mapping $D_1 H$ and so the rotational indices of the two curves coincide. In fact, the converse holds, a fact which we state without proof:

Proposition 2.6 *(Whitney and Grauenstein): If c_1 and c_2 are closed curves with the same rotational index, then they are isotopic.*

We now consider one of the most famous of geometrical inequalities—the so-called **isoperimetric inequality.** Let $c : [a, b] \to \mathbf{R}^2$ be a simple, closed curve. Then it divides the plane into two open, connected regions—its interior and exterior (this is a special case of the Jordan curve theorem). We shall only require the following formula for the area of the interior:

$$A = \int_a^b c_1(t)\dot{c}_2(t)dt = -\int_a^b c_2(t)\dot{c}_1(t)dt$$

$$= \int_a^b \frac{1}{2}[c_1(t)\dot{c}_2(t) - c_2(t)\dot{c}_1(t)]dt.$$

Lemma 2.7 *Let f be a smooth 2π-periodic function on the line. Then if $\int_0^{2\pi} f(t)dt = 0$, f satisfies the following inequality:*

$$\int_0^{2\pi} |f(t)|^2 dt \leq \int_0^{2\pi} |f'(t)|^2 dt.$$

There is equality if and only it f has the form $f(t) = a\cos t + b\sin t$ for suitable real numbers a and b.

Proof. We consider the Fourier series representation

$$\sum_{n=1}^{\infty} a_n \cos nt + b_n \sin nt$$

of f. The derivative of f has Fourier series

$$\sum_{n=1}^{\infty} nb_n \cos nt - na_n \sin nt.$$

We have the formulae

$$\int_0^{2\pi} |f(t)|^2 dt = \sum_{n=1}^{\infty} (a_n^2 + b_n^2)$$

$$\int_0^{2\pi} |f'(t)|^2 dt = \sum_{n=1}^{\infty} n(a_n^2 + b_n^2)$$

from which the result easily follows.

■

Proposition 2.8 *Let c be a simple, closed curve in the plane. Then we have the inequality $4\pi A \leq L^2$ where L is the length of the curve and A is the area of its interior.*

PROOF. We simplify the notation by assuming that the curve has length 2π and is parametrised by arc-length. Furthermore we can assume that the y-axis passes through the centroid of the figure formed by the curve i.e. that $\int_0^{2\pi} c_1(t)dt = 0$. Then $A = \int_0^{2\pi} c_1(t)\dot{c}_2(t)dt$ and so

$$2\pi - 2A = \int_0^{2\pi} [\dot{c}_1(t)^2 + \dot{c}_2(t)^2 - 2c_1(t)\dot{c}_2(t)]dt$$

$$= \int_0^{2\pi} [\dot{c}_1(t)^2 - c_1(t)^2]dt + \int_0^{2\pi} (c_1(t) - \dot{c}_2(t))^2 dt$$

and both terms of the right hand side are non-negative. Hence $L = 2\pi \geq 2A$ i.e. $L^2 = 4\pi^2 \geq 4\pi A$.

■

We remark that by examining in detail what happens when one has equality, one can show that this implies that the curve is a circle.

We now turn to results on convex curves. These are defined to be simple, closed curves which lie on the same side of their tangents i.e. are such that for each $s_0 \in [0, L]$, the sign of $\big(\gamma(s) - \gamma(s_0)|\mathbf{N}(s_0)\big)$ is constant. We can characterise them in terms of the curvature as follows:

Proposition 2.9 *Let γ be a simple, closed curve. Then γ is convex if and only if κ has constant sign.*

PROOF. We suppose firstly that γ is convex. Choose $\tilde{\theta} : [0, L] \to \mathbf{R}$ so that $p \circ \tilde{\theta} = \mathbf{T}$. By Taylor's theorem,

$$\gamma(s) = \gamma(s_0) + (s - s_0)\mathbf{T}(s_0) + (s - s_0)^2\kappa(s_0)\mathbf{N}(s_0) + R(s)$$

where the remainder term R satisfies the growth condition

$$\lim_{s \to s_0} \frac{R(s)}{(s - s_0)^2} = 0.$$

Then

$$(\gamma(s) - \gamma(s_0)|\mathbf{N}(s_0)) = (s - s_0)^2\kappa(s_0) + (R(s)|\mathbf{N}(s_0)).$$

Hence if the left hand side has constant sign, then so also has κ.

On the other hand, if γ is not convex, then there is an $s_0 \in [0, L]$ so that $f(s) \leq 0$ infinitely close to s_0 in $]s_0 - \epsilon, s_0[$ and $f(s) \geq 0$ infinitely close to s_0 in $]s_0, s_0 + \epsilon[$ where $f(s) = (\gamma(s) - \gamma(s_0)|\mathbf{N}(s))$. This implies that $f'(s_0) = 0$. Suppose that f attains its maximum resp. minimum at s_1 and s_2. Then its derivative

$f'(s)$ vanishes at the three points s_0, s_1, s_2. Hence there are two values of s for which $\mathbf{T}(s)$ coincide (since $\mathbf{T}(s_0)$, $\mathbf{T}(s_1)$ and $\mathbf{T}(s_2)$ are all unit vectors which are perpendicular to $\mathbf{N}(s_0)$). We can suppose, without loss of generality, that the parametrisation is so chosen that the two points where this happens are 0 and s'. If we assume that the curvature is non-negative, then there are positive integers k and k' so that

$$\tilde{\theta}(L) - \tilde{\theta}(s') = 2\pi k' \quad \text{and} \quad \tilde{\theta}(s') - \tilde{\theta}(0) = 2\pi k.$$

Then $k + k' = 1$ and this is clearly impossible.

We remark that since the integral of the curvature from a to b is $\pm 2\pi$, either of the above conditions is equivalent to the fact that $\int_a^b |\kappa(s)| ds = 2\pi$.

We mention without proof that the above characterisations of convexity are also equivalent to the more familiar description that the curve is the boundary of a region in the plane which is convex in the classical sense.

We now discuss **vertices** of curves. These are points where the derivative of the curvature function vanishes. In general, the curvature attains a local maximum or minimum there (although this need not always be the case). A non-circular ellipse, for example, has four vertices – the ends of the major axes. The ends of the longer of the two major axes are points of maximum curvature. The next result – the famous **four vertex theorem** – shows that the behaviour of the ellipse is in a certain sense typical:

Proposition 2.10 *Let γ be a smooth, simple, closed curve. Then it has at least four vertices.*

We shall prove this result under the further assumption that the curve is convex. PROOF. κ has a maximum and a minimum which are distinct unless κ is constant, in which case the curve is a circle. Then every point is a vertex. Hence we always have at least two vertices. We shall suppose that there are only two and obtain a contradiction. The two vertices divide the curve into two parts, on one of which κ' is non-negative and on one of which it is non-positive. We can assume that the parametrisation is so chosen that κ assumes its minimum at 0 and its maximum at $s' \in [0, L]$. By rotating the coordinate axes, we can suppose in addition that $\gamma(0)$ and $\gamma(s')$ lie on the x-axis. Now the latter meets the curve at these two points only and so we can arrange for the part of the curve with $\kappa' \leq 0$ to be the part above the x-axis. Then $\kappa'(s)\gamma(s) \leq 0$ for each s. By the Frenet formula $\gamma_1'' = -\kappa\gamma_2'$ and so we have the inequality

$$0 \geq \int_0^L \kappa'\gamma_2 ds = -\int_0^L \kappa\gamma_2' ds$$

by integration by parts and the latter is

$$\int_0^L \gamma_1'' ds = \mathbf{T}_1(L) - \mathbf{T}_1(0) = 0.$$

Hence, since the integrand $\kappa'\gamma_2$ has constant sign, we have that it vanishes identically, which is obviously impossible. Thus, γ has at least three vertices. If it only has three, then two of them would divide the curve into two parts, on one of which the derivative of the curvature is non-negative and on the other of which it is non-positive. But the above argument shows that this is impossible. Hence the curve has at least four vertices.

■

3 CURVES IN SPACE

In this chapter, we discuss three-dimensional curves. We shall confine ourselves to the analogues of moving frames, curvature and approximating circles. The treatment will be similar to that of the second chapter but the extra dimension leads to the introduction of a new parameter - torsion - which describes the tendency of the curve to twist out of the best approximating plane on which it lies. Also the circle of curvature will be replaced by a sphere.

Definition: A C^r**-parametrised curve** in $\mathbf{R}^3$ is a C^r mapping $c : I \to \mathbf{R}^3$ where I is an interval in $\mathbf{R}$. The curve is **regular** if $\dot{c}(t) \neq 0$ for each $t \in I$.

As in the case of plane curve, we identify c and $c \circ \phi^{-1}$ where $\phi : I \to J$ is a reparametrisation. In particular, we can define a particular reparametrisation ϕ given by exactly the same formula as in the planar case (so that its derivative $\dot{\phi}$ is $|\dot{c}|$). This reparametrisation induces **parametrisation by arc-length**.

Example: The functions

$$c : t \mapsto (a\cos t, a\sin t, bt) \quad (t \in]-\pi, \pi[)$$

and

$$c : u \mapsto (a\frac{1-u^2}{1+u^2}, \frac{2au}{1+u^2}, 2b\arctan u) \quad (u \in \mathbf{R})$$

are parametrisations of the same curves (one turn of a helix) in space.

If $\gamma : I \to \mathbf{R}^3$ is a C^r-curve, parametrised by arc-length, and s_0 is a point in I with $\gamma''(s_0) \neq 0$, then the **osculating plane** to γ at s_0 is the plane through $\gamma(s_0)$, parallel to $\gamma'(s_0)$ and $\gamma''(s_0)$. Hence it has the equation

$$(x - \gamma(s_0)|\gamma'(s_0) \times \gamma''(s_0)) = 0.$$

If we substitute the derivatives of c according to the formulae

$$c(t) = \gamma(s), \quad \dot{c}(t) = \gamma'(s)\dot{\phi}(t),$$

$$\ddot{c}(t) = \gamma''(s)\dot{\phi}(t)^2 + \gamma'(s)\ddot{\phi}(t)$$

where $s = \phi(t)$, then the equation takes the form

$$(x - c(t_0)|\dot{c}(t_0) \times \ddot{c}(t_0)) = 0$$

or

$$\det \begin{bmatrix} \xi_1 - c_1(t_0) & \xi_2 - c_2(t_0) \\ \xi_3 - c_3(t_0) & \\ \dot{c}_1(t_0) & \dot{c}_2(t_0) \\ \dot{c}_3(t_0) & \\ \ddot{c}_1(t_0) & \ddot{c}_2(t_0) \\ \ddot{c}_3(t_0) & \end{bmatrix} = 0.$$

The definition is motivated by the fact that when $\gamma''(s_0) \neq 0$ then there is a neighbourhood U of s_0 so that $\gamma(s_0), \gamma(s_1), \gamma(s_2)$ are not collinear when s_0, s_1, s_2 are distinct points in U. The the plane through $\gamma(s_0), \gamma(s_1), \gamma(s_2)$ is then well-defined and tends to the above one as s_1 and s_2 tend to s_0.

Examples: The osculating plane to the curve

$$c : t \mapsto (t, t^2, t^3)$$

at t_0 has equation

$$\det \begin{bmatrix} \xi_1 - t_0 & \xi_2 - t_0^2 & \xi_3 - t_0^3 \\ 1 & 2t_0 & 3t_0^2 \\ 0 & 2 & 6t_0 \end{bmatrix} = 0$$

i.e. $3t_0^2\xi_1 - 3t_0\xi_2 + \xi_3 - t_0^3 = 0$.

The example of the curve

$$c : t \mapsto \begin{cases} (e^{-\frac{1}{t^2}}, t, 0) & (t \leq 0) \\ (0, t, e^{-\frac{1}{t^2}}) & (t \geq 0) \\ (0, 0, 0) & (t = 0) \end{cases}$$

at the point 0 shows that some condition on γ is necessary to ensure the existence of an osculating plane.

Definition: If $s_0 \in I$, the **tangent vector** to the curve γ at s_0 is the (unit) vector $\mathbf{T}(s_0) = \gamma'(s_0)$. The **curvature** $\kappa(s_0)$ is defined to be $|\mathbf{T}'(s_0)| = |\gamma''(s_0)|$. The **normal plane** to γ at s_0 is the plane through $\gamma(s_0)$, perpendicular to the tangent vector i.e. it has the equation

$$(x - \gamma(s_0) | \mathbf{T}(s_0)) = 0.$$

The **principal normal** to γ at s_0 is the unit vector

$$\mathbf{N}(s_0) = \frac{\mathbf{T}'(s_0)}{|\mathbf{T}'(s_0)|} = \frac{\gamma''(s_0)}{|\gamma''(s_0)|}.$$

The **binormal** $\mathbf{B}(s_0)$ at s_0 is the vector

$$\mathbf{B}(s_0) = \mathbf{T}(s_0) \times \mathbf{N}(s_0) = \frac{\gamma'(s_0) \times \gamma''(s_0)}{|\gamma''(s_0)|}.$$

Then the triple $(\mathbf{T}, \mathbf{N}, \mathbf{B})$ at (s_0) forms a positively oriented orthonormal basis for $\mathbf{R}^3$—called the **moving frame** of γ. Note that the osculating plane is the plane through $\gamma(s_0)$, parallel to $\mathbf{T}$ and $\mathbf{N}$. Similarly, we call the plane parallel to $\mathbf{N}$ and $\mathbf{B}$ the **normal plane** and that parallel to $\mathbf{T}$ and $\mathbf{B}$ the **rectifying plane.**

The curvature describes how the curve is turning within the osculating plane. We now introduce a scalar function which indicates how it is twisting out of the latter plane. It is called the **torsion** and is defined as follows: since the norm of the binormal $\mathbf{B}$ is constant (and in fact one), we know that $\mathbf{B}'$ and $\mathbf{B}$ are perpendicular. For $0 = (\mathbf{B}|\mathbf{B})' = 2(\mathbf{B}|\mathbf{B}'))$. Also $\mathbf{B}$ and $\mathbf{T}$ are perpendicular and if we differentiate the corresponding relationship $(\mathbf{B}|\mathbf{T}) = 0$, we get

$$(\mathbf{B}'|\mathbf{T}) = -(\mathbf{B}|\mathbf{T}') = -\kappa(\mathbf{B}|\mathbf{N}) = 0.$$

Hence $\mathbf{B}'$ is a multiple of $\mathbf{N}$ and so we can define the torsion to be the scalar function τ so that $\mathbf{B}' = -\tau\mathbf{N}$.

We have the following explicit formula for τ: $\tau = (-\mathbf{N}|\mathbf{B}') = (-\mathbf{N}|(\mathbf{T}\times\mathbf{N})')$
$= (-\mathbf{N}|\mathbf{T}\times\mathbf{N}') + (\mathbf{N}|\mathbf{T}'\times\mathbf{N})$
$= \left(-\frac{\gamma''}{\kappa}|\gamma'\times\left(\frac{\gamma''}{\kappa}\right)'\right)$
$= \left(-\frac{\gamma''}{\kappa}|\gamma'\times\left(\frac{\kappa\gamma'''-\kappa'\gamma''}{\kappa^2}\right)\right)$
$= \frac{-1}{\kappa^2}\big(\gamma''|\gamma'\times\gamma'''\big) = \frac{1}{\kappa^2}\big(\gamma'|\gamma''\times\gamma'''\big)$. For an arbitrary parametrisation c substitution in the above leads to the formula

$$\frac{(\dot{c}|\ddot{c}\times\dddot{c})}{|\dot{c}\times\ddot{c}|^2}$$

for the torsion.

Corresponding to the formula for the derivatives of $\mathbf{T}$ and $\mathbf{N}$ for planar curves, we now have the following expressions for the derivatives of $\mathbf{T}$, $\mathbf{N}$ and $\mathbf{B}$:

$$\begin{aligned}\mathbf{T}' &= \kappa\mathbf{N}\\ \mathbf{N}' &= -\kappa\mathbf{T}+\tau\mathbf{B}\\ \mathbf{B}' &= -\tau\mathbf{N}.\end{aligned}$$

which we can write in matrix form:

$$\begin{bmatrix}\mathbf{T}\\ \mathbf{N}\\ \mathbf{B}\end{bmatrix}' = \begin{bmatrix}0 & \kappa & 0\\ -\kappa & 0 & \tau\\ 0 & -\tau & 0\end{bmatrix}\begin{bmatrix}\mathbf{T}\\ \mathbf{N}\\ \mathbf{B}\end{bmatrix}$$

(These are called the **Serret-Frenet formulae**).

Proof. The first and last lines are the definitions of the curvature and the torsion. In order to verify the middle line, note that $\mathbf{N}'$ and $\mathbf{N}$ are perpendicular and so the derivative of $\mathbf{N}$ has the form $\alpha\mathbf{T}+\beta\mathbf{B}$ for scalar functions α and β which can be calculated as follows: since $(\mathbf{N}|\mathbf{T}) = 0, (\mathbf{N}'|\mathbf{T}) = -(\mathbf{N}|\mathbf{T}') = -\kappa$. Since $(\mathbf{T}|\mathbf{B}) = 0, (\mathbf{N}'|\mathbf{B}) = -(\mathbf{N}|\mathbf{B}') = \tau$.

■

Now if c is an arbitrary parametrisation of the curve and we define $\mathbf{T}_{c(t)}$ to be $\mathbf{T}_{\gamma(s)}$ etc., then we have

$$\begin{aligned}\mathbf{T}_c &= \kappa_c\dot{\phi}\mathbf{N}_c \\ \mathbf{N}_c &= -\kappa_c\dot{\phi}\mathbf{T}_c + \tau_c\dot{\phi}\mathbf{B} \\ \mathbf{B}_c &= -\tau_c\dot{\phi}\mathbf{N}_c.\end{aligned}$$

Our definition of the moving frame has the disadvantage that we can only apply it to curves for which the second derivative of the parametrisation never vanishes. Of course some restriction is necessary as the example considered above shows but the above is unnecessarily restrictive. For example, we cannot assert that the curvature and torsion of a straight line are zero as they clearly should be. Hence we extend our definition as follows:

A **moving frame** for a curve with parametrisation γ is a triple $(\mathbf{T}, \mathbf{N}, \mathbf{B})$ of continuous functions from the interval I of definition into $\mathbf{R}^3$ so that
a) for each $s \in I, (\mathbf{T}(s), \mathbf{N}(s), \mathbf{B}(s))$ is a positively orientated orthonormal basis for three-dimensional space:
b) $\mathbf{T}$ is the derivative of γ;
c) there are continuous functions f, g so that $\mathbf{T}, \mathbf{N}, \mathbf{B}$ satisfy the following conditions:

$$\begin{aligned}\mathbf{T}' &= f\mathbf{N} \\ \mathbf{N}' &= -f\mathbf{T} + g\mathbf{B} \\ \mathbf{B}' &= -g\mathbf{N}.\end{aligned}$$

If a curve has a moving frame, then we define the curvature κ to be the function f and the torsion τ to be the function g. As we have seen, this definition does not contradict the original one for points where the second derivative of the parametrisation does not vanish.

Example: Consider the curve $c : t \mapsto (t, t^3, 0)$. In this case the second derivative vanishes at the origin and so, under our original definition, the torsion and curvature are not defined there. However, it is clear that the curve has a moving frame which at the origin coincides with the canonical basis.

Example: Calculate the curvature and torsion function of the curve $c : t \mapsto (t, t^2, t^3)$.
We have

$$\dot{c}(t) = (1, 2t, 3t^2)$$

and so

$$\dot{\phi}(t) = (1 + 4t^2 + 9t^4)^{\frac{1}{2}}.$$

Differentiating the equation

$$\dot{\phi}(t)\mathbf{T}(t) = (1, 2t, 3t^2)$$

we get

$$\ddot{\phi}(t)\mathbf{T}(t) + \dot{\phi}^2(t)\kappa(t)\mathbf{N}(t) = (0, 2, 6t).$$

Taking the cross product of the two equations gives

$$\dot{\phi}^3\kappa\mathbf{B} = 2(3t^2, -3t, 1)$$

and so

$$\kappa^2 = 4\frac{(9t^4 + 9t^2 + 1)}{(9t^4 + 4t^2 + 1)^3}.$$

If we differentiate the second last equation, we get

$$\dot{\phi}^3\kappa\mathbf{B} - 4\dot{\phi}\kappa\tau\mathbf{N} = 6(2t, -1, 0).$$

Taking scalar products, we get $-\dot{\phi}^6\kappa^2\tau = -12$ and so $\tau = 3(9t^4 + 9t^2 + 1)^{-1}$.

In order to emphasise the geometrical significance of the curvature and torsion function, we consider the Taylor expansion of the parametrising function γ. Here we assume that the point we are interested in has parameter 0 and that the derivative of the parametrisation there is non-zero. Then

$$\gamma(s) = \gamma(0) + s\gamma'(0) + \frac{s^2}{2!}\gamma''(0) + \frac{s^3}{3!}\gamma'''(0) + \frac{s^4}{4!}\gamma^{(4)}(0)$$

plus a remainder term which we shall ignore.

Using the Serret-Frenet formulae we can rewrite this in the form

Hence if we consider the coordinate of the curve with respect to the moving frame i.e. if we define functions X, Y, Z by the equations

$$\gamma(s) - \gamma(0) = X(s)\mathbf{T}(0) + Y(s)\mathbf{N}(0) + Z(0)\mathbf{B}(0)$$

then

A useful aid for visualising the geometrical significance of the curvature and torsion is as follows: if γ is a curve, then we have three corresponding curves

$$t \mapsto \mathbf{T}_\gamma(t),\ t \mapsto \mathbf{N}_\gamma(t)$$

$$t \mapsto \mathbf{B}_\gamma(t)$$

which lie on the unit sphere S^2 (these curves need not be regular). They are called the **spherical indicatrix** of the tangent resp. the normal resp. the binormal. If we denote the corresponding length fuctions by $\phi_\mathbf{T}, \phi_\mathbf{N}, \phi_\mathbf{B}$ resp., then a simple calculation shows that κ and τ are the absolute values of the rates of change of $\phi_\mathbf{T}$ and $\phi_\mathbf{B}$.

Since the torsion measures the tendency of the curve to twist away from the osculating plane, the following result is not surprising:

Proposition 3.1 *A curve is planar (i.e. lies in a plane) if and only if the torsion function τ vanishes.*

PROOF. Suppose that the torsion vanishes. Then $\mathbf{B}' = 0$ i.e. there is a constant vector c so that $\mathbf{B} = c$. Hence $(\gamma'|c) = 0$ i.e. the function $(\gamma|c)$ is constant. In other words, if s_0 is a fixed point in I, then $(\gamma(s) - \gamma(s_0)|c) = 0$. But this means that γ lies on the plane through $\gamma(s_0)$ perpendicular to c.
On the other hand, if γ lies on a plane, there is a fixed unit vector c and a point s_0 in I so that $(\gamma(s) - \gamma(s_0)|c) = 0$. Then we can reverse the above reasoning to show that the torsion vanishes.

■

The intrinsic equation of a curve: Analogue to the two-dimensional case, knowledge of the torsion and curvature determines a curve up to its position in space. Once again, the proof is an application of existence and uniqueness theorems for ordinary differential equations.

Proposition 3.2 *Let f and g be continuous functions on the interval $[0, L]$ and suppose that the function f is non-negative. Then there is a curve with parametrisation γ (defined on $[0, L]$) which has f as its curvature function and g as its torsion. Further γ is unique up to a direct isometry of space.*

We conclude this chapter with an alternative approach to the topics of curvature, torsion etc. The osculating plane and the sphere of curvature are characterised by the high level of contact that they enjoy with the curve. This notion is made precise in the following definition:

Definition: Consider the point p on the surface $S = \{x : f(x) = 0\}$ and the curve γ where $\gamma(s_0) = p$. Then S and γ have **k-point contact** (or **k-fold contact**) at s_0 if the composed function $\phi = f \circ \gamma$ is such that

$$\phi(s_0) = \phi'(s_0) = \cdots = \phi^{(k-1)}(s_0) = 0, \quad \phi^{(k)}(s_0) \neq 0$$

i.e.

$$\phi(s) = (s - s_0)^{k-1}\tilde{\phi}(s)$$

where $\tilde{\phi}(s_0) \neq 0$.

We use this definition to determine those planes, spheres etc. which have best possible contact with a given curve. For the sake of simplicity, we begin with curves in the plane:

The height function: Consider the function $\phi(s) = (\gamma(s)|v)$ where v is a unit vector. Up to a constant, this is the function which describes the contact of the curve with the line $(x - \gamma(s_0)|v) = 0$. In this case $\phi'(s) = (\mathbf{T}(s)|v)$ and we see that this vanishes provided that v is perpendicular to the tangent vector i.e. v is (up to sign) the normal vector. In other words, a line has at least 2-fold contact if and only if it is the tangent. The reader can check that it has higher order contact if and only if it is the tangent at a point of inflection.

The distance-squared function: This is the function

$$\phi(s) = (\gamma(s) - x|\gamma(s) - x).$$

Then

$$\phi'(s) = 2(\gamma(s) - x|\mathbf{T}(s)),$$
$$\phi''(s) = 2(1 + \kappa(s)(\gamma(s) - x|\mathbf{N}(s))),$$
$$\phi'''(s) = 2[\kappa'(s)(\gamma(s) - x|\mathbf{N}(s)) - \kappa^2(s)(\gamma(s) - x|\mathbf{T}(s))].$$

Hence the circle with centre x through $\gamma(s_0)$ has at least two point contact if and only if x is on the normal. It has at least three point contact if x is the centre of curvature and it has four point contact if x is the centre of curvature at a vertex.

If we consider the corresponding functions in space, we get:

The height function (which determines contact with planes): $\phi(s) = (\gamma(s)|v)$ and $\phi'(s_0) = 0$ if and only if the vector v lies in the normal plane. The first two derivatives vanish if and only if v (up to sign) is the binormal. We are tacitly assuming that the curvature at the given point does not vanish).

The distance-squared function (determines contact by spheres):

$$\phi(s) = (\gamma(s) - x|\gamma(s) - x)$$
$$\phi'(s) = 2(\gamma(s) - x|\mathbf{T}(s))$$
$$\phi''(s) = 2(1 + \kappa(s)(\gamma(s) - x|\mathbf{N}(s)))$$
$$\phi'''(s) = 2[\kappa'(s)(\gamma(s) - x|\mathbf{N}(s)) + \kappa(s)(\gamma(s) - x| - \kappa(s)\mathbf{T}(s) + \tau(s)\mathbf{B}(s))].$$

Hence we have two-fold contact if x is in the rectifying plane, three point contact if

$$x = \gamma(s) + \frac{\mathbf{N}(s)}{\kappa(s)} + \lambda\mathbf{B}(s)$$

for some λ and four point contact if

$$x = \gamma(s) + \frac{\mathbf{N}(s)}{\kappa(s)} - \frac{\kappa'(s)}{\kappa^2(s)\tau(s)}\mathbf{B}(s).$$

(provided that $\tau(s) \neq 0$).

4 CONSTRUCTION OF CURVES

In this section we shall describe various methods of obtaining new curves from old ones. The emphasis will be on translating some natural geometrical constructions (such as rolling circles along curves or reflecting curves in curvilinear mirrors) into analytic terms. This provides a certain unifying thread in the construction or classification of concrete curves. Among the methods which we shall discuss are the following: conchoids, involutes, evolutes, strophoids, glissettes, roulettes, envelopes, pedal curves, group actions and actions on space.

Evolutes: Suppose that γ is a curve with non-vanishing curvature. Then the curve

$$E_\gamma(t) = \gamma(t) + \rho(t)\mathbf{N}(t)$$

(i.e. the locus of the centres of curvature of c) is called the **evolute** of γ. Of course E_γ need not be regular (e.g. if γ is a circle, it reduces to a point). This is a general phenomenon—such derived curves can have singularities which often correspond to significant geometrical properties of the original curve. For example, the evolute has a singularity exactly when $\rho'(t) = 0$ i.e. at a vertex. For we calculate the tangent to the evolute as follows: differentiating the equation

$$E_\gamma(t) = \gamma(t) + \rho(t)\mathbf{N}(t)$$

we get

$$E'_\gamma(t) = \mathbf{T}(t) + \rho'(t)\mathbf{N}(t) - \rho(t)\kappa(t)\mathbf{T}(t) = \rho'(t)\mathbf{N}(t).$$

Hence the curve fails to be regular at a point where the derivative of ρ vanishes. But the derivative of ρ vanishes precisely at the vertices of the curve (since $\rho' = -\frac{\kappa'}{\kappa^2}$). The typical example of this is the parabola where the evolute has a cusp corresponding to the vertex.

We note some simple properties of the evolute which follow from the definition:
1) the normals to the curve are tangential to the evolutes. For the normal at $\gamma(t)$ meets the curve at $E_\gamma(t)$ and the tangent to E_γ is parallel to $\mathbf{N}(t)$ as we calculated above.
2) if the curvature function κ is strictly monotone, then the equation of the evolute can be written in the form

$$E_\gamma(t) = \left(\int_{t_o}^{t} \cos\theta(\sigma)d\sigma - \frac{1}{\theta'(t)}\sin\theta(t), \int_{t_0}^{t} \sin\theta(\sigma)d\sigma + \frac{1}{\theta'(t)}\cos\theta(t)\right)$$

where $\mathbf{T}(t) = (\cos\theta(t), \sin\theta(t))$.

Examples of evolutes: The evolute of the parabola is a semi-cubic parabola (with cusp corresponding to the vertex of the parabola as we have seen). The evolute of an ellipse is an asteroid, of a hyperbola a Lamé curve, of an epicycloid another epicycloid and of a hypocycloid again a hypocycloid.

Involutes: An involute of a curve c is one of the form

$$I_c(t) = c(t) + (a - \int_{t_O}^{t} |\dot{c}(u)|du)\mathbf{T}_c(t)$$

where a is a fixed real number (we say **an** involute since the curve is dependent on the choice of a). It is the locus of one endpoint of a piece of thread (the other end of which is attached to the curve) as the thread is laid along the curve.

In the case where the curve is parametrised by arc-length these equations take on the simplified form

$$I_\gamma(t) = \gamma(t) + (t_0 - t)\mathbf{T}(t).$$

Note that we lose one degree of differentiability when we form the involute. For example the involute of a C^3-curve is C^2. We note some geometrical properties of the involute:
1) the involute is orthogonal to the tangents of the original curve at the corresponding point. For $\dfrac{dI_\gamma}{dt} = -\kappa(t)\mathbf{N}(t)$.

2) the evolute of the involute is the original curve. Hence the examples of evolutes provide examples of involutes.

The pedal of a curve is the locus of the feet of the perpendiculars from a fixed point to the tangents of the curve. If we take the origin as fixed point then the pedal P_γ has the equation

$$P_\gamma(t) = (\gamma(t)|\mathbf{N}(t))\mathbf{N}(t).$$

Its derivative is

$$\dot{P}_\gamma(t) = \kappa(t)[(\gamma(t)|\mathbf{T}(t))\mathbf{N}(t) + (\gamma(t)|\mathbf{N}(t))\mathbf{T}(t)].$$

Hence if the curve does not pass through the origin the pedal is regular except at points which correspond to points of inflection of γ.

Examples of pedal curves: The pedal to the parabola (with pedal point the vertex) is the cissoid of Diocles. The pedal to the circle with an arbitrary point on the circumference as pole is the cardioid. A pedal to the circle, with any other point as pole, is a Limaçon.

Mirror images: We consider a curve c and a point O which does not lie on c. We roll a mirror along c and consider the curve traced by the image of A in these mirrors. This produces the curve

$$M_c(t) = x_0 - 2(x_0 - c(t)|\mathbf{N}(t))\mathbf{N}(t)$$

which reduces to the form

$$M_c(t) = -2(c(t)|\mathbf{N}(t))\mathbf{N}(t)$$

in the case where O is the origin. This curve is called the **orthotomic** of c and is related to the pedal curve as one sees directly from the above formula.

Parallel curves: The family

$$c_r(t) = c(t) + r\mathbf{N}(t)$$

of curves (where r is an additional parameter) is called the family of parallels to c. Note that c_r has a singularity at a point t_0 where $\kappa(t_0) = \dfrac{1}{r}$ i.e. at the points where the parallel curve intersects the evolute. The parallel then generally has a cusp at such points. The typical example is the family of parallels of a parabola (see figure).

We remark that the family of involutes of a given curve form a parallel family.

The following are simple properties of the parallel curves:

a) the normal vector $\mathbf{N}(t)$ to the original curve is also normal to the parallel curves (at the same value of the parameter t).

b) the relationship between c and c_r is symmetric i.e. c is also a parallel to c_r.

c) the tangents to c and c_r at corresponding points are parallel.

Roulettes: We suppose that we have two curves γ_1 and γ_2 both parametrised by arc-length and we construct a new curve which is called a roulette as follows: if $s \in I$, there is an angle $\theta(s)$ so that $D_{\theta(s)}$ maps $\mathbf{T}_{\gamma_1}(s)$ onto $\mathbf{T}_{\gamma_2}(s)$ (the existence of a smooth function θ with this property follows from the result of Chapter 2). The isometry

$$U(s) = T_{\left(\gamma_2(s)-\gamma_1(s)\right)} \circ D_{\theta(s)}$$

maps $\gamma_1(s)$ onto $\gamma_2(s)$ and revolves the first curve around the axis $\gamma_1(s)$ until its tangent lies along that of γ_2. In other words, the family of isometries $U(s)$ describes the motion of rolling the first curve along the second one. The image of a point under this motion i.e. a curve of the form $R_{\gamma_1,\gamma_2}(t) = U(t)x_0$ for some fixed x_0 is called a roulette.

Examples of roulettes are cycloids, hypocycloids, epicycloids, Cardano's circles, planetary paths.

Conchoids: We start with a curve c, a point A not on c and a constant k. The conchoid of c with respect to A is the locus of the point P on the line AQ which is such that the length $|QP|$ is equal to the constant k as Q varies on the curve. If A has coordinates x_0 then the parametrisation of the conchoid is

$$C_c(t) = c(t) + \frac{k(c(t) - x_0)}{|c(t) - x_0|}.$$

The classical example is the conchoid of Nicodemes which is the conchoid of a point with respect to a straight line.

Cissoids: If c_1 and c_2 are curves (for reasons which will soon be apparent it is not convenient to assume here that they are parametrised by arc-length) and A is a fixed point which does not lie on either of them, the cissoid of the two curves with respect to A is the locus of a point P which moves as follows: we choose a point Q on the first curve and let the line AO (produced if necessary) meet the second curve at R. (We are assuming that the latter line meets this curve at precisely one point or, if it meets it at several points, that there is a natural choice of one of these as R). P is then defined to be the point on the line AQ which is such that $|AP| = |QR|$.

Analytically this means that if we assign the coordinates x_0 to A, then parametrisations can be chosen in such a way that the three points x_0 $c_1(t)$ and $c_2(t)$ are always collinear. The cissoid is then the curve with parametrisation

$$C_{c_1,c_2}(t) = x_0 + c_1(t) - c_2(t).$$

Example: The cissoid of Diocles is the cissoid of a circle and its tangent with respect to the point of the circle directly opposite to the point of contact (see diagram).

Strophoids: A curve c and two fixed points O and A are given. If Q lies on c and P is the point on the line OQ so that $|QP| = |AQ|$, the locus of P as Q traces out c is called the strophoid of c with respect to O and A. It has parametrisation

$$S_c(t) = c(t) + |c(t) - x_A| \frac{c(t) - x_0}{|c(t) - x_0|}.$$

The classical example is the strophoid of a straight line with respect to a pole O not on the line and with fixed point A the foot of the perpendicular from O to the line. If we choose for A a point not on this perpendicular, then we get a so-called oblique strophoid.

Transformations of curves: If c is a curve in an open subset U of the plane and ϕ is a suitable mapping (for example, a diffeomorphism) from U into the plane, then the image $\phi(c)$ and pre-image $\phi^{-1}(c)$ of c are also curves. This provides a method of obtaining new curves from simpler ones. One of the commonest such transformations used in elementary geometry is the mapping

$$\phi : x \mapsto \frac{x}{|x|^2}$$

of inversion in the unit circle (this is a diffeomorphism of the punctured plane). In complex coordinates, this is the mapping $z \mapsto 1/\bar{z}$. We remark here that this function is not holomorphic but it is anti-holomorphic so that it preserves angles (howbeit with a reversal of signs).

Examples: If we invert a central conic in a circle with centre at a focus, the the result is a limaçon. If we invert a central conic in a circle with the same centre then we get a Cassinian. The inversion of a hyperbola in a circle with the same centre is the leminiscate of Bernoulli. The inversion of a hyperbola in a circle with centre at a vertex is a strophoid.

We can generalise the above method of obtaining new curves in several ways. For example we can consider functions ϕ of two variables which can then be applied to two curves c_1 and c_2 to give a new curve with parametrisation

$$c_3 : t \mapsto \phi(c_1(t), c_2(t)).$$

Examples of such constructions are those of the strophoids, the cissoids and the conchoids.

Still more generally, one can consider transformations which involve derivatives as in the case of the constructions we used to produce roulettes and pedal curves.

Envelopes: One of the most attractive methods of constructing curves is by means of envelopes of families of curves. We begin with the concrete case of the

family of circles of radius 1 with centres on the x-axis. Classically, the envelope is defined to be a curve (or curves) which touch each member of the given family tangentially. In the above case, it is clear that the envelope is the pair of straight lines with equations $\xi_2 = \pm 1$. We can describe this geometrically as follows. The above family can be specified by the equations

$$(\xi_1 - t)^2 + \xi_2^2 - 1 = 0$$

where t is now the parameter which specifies which particular member of the family we are describing. Now the above equation is that of a surface in three dimensional space (with coordinates (ξ_1, ξ_2, t)) - an oblique cylinder. The circles are the projections onto the (ξ_1, ξ_2)-plane of its cross-sections with horizontal planes. The two enveloping curves are the apparent contours of this surface as seen from below. If we denote by F the function

$$F(x,t) = (\xi_1 - t)^2 + \xi_2^2 - 1,$$

then this set is characterised as follows:

$$E_F = \{x \in \mathbf{R}^2 : \quad \text{there is a} \quad t \quad \text{so that} \quad F(x,t) = D_t F(x,t) = 0\}.$$

The general case is described in the following definition:

Definition: Let f be a smooth function on an open subset U of the plane. Then a (one-dimensional) **unfolding** of f is a smooth function F on $U \times \mathbf{R}$ so that $F(x,0) = f(x)$ for all x in U. Thus the above function F is an unfolding of $f : x \mapsto \xi_1^2 + \xi_2^2 - 1$. As a second example

$$F : (x,t) \mapsto 2t^3 + t(1 - 2\xi_2) - \xi_1$$

is an unfolding of the function $x \mapsto \xi_1$ of projection onto the first coordinate.

An unfolding of the general form described in the definition embeds the initial curve defined implicitly by the equation $f(x) = 0$ into the family c_t, where $c_t = \{x : F(x,t) = 0\}$. The **envelope** of a family of curves defined by an unfolding is defined to be the set

$$E_F = \{x : \quad \text{there is a} \quad t \quad \text{with} \quad F(x,t) = D_t F(x,t) = 0\}.$$

In general (but not always), this will be a curve which is tangential to each c_t. We shall not discuss in detail the condition required to ensure that the above set really is a curve.

Examples:

I. Consider the unfolding

$$F(x,t) = (x - \gamma(t) | x - \gamma(t)) - r^2$$

where γ is a given curve. This corresponds to the family of circles with centres on the curve with the same radius r. Then

$$E_F = \{x : x = \gamma(t) \pm r\mathbf{N}(t) \quad \text{for some} \quad t\}.$$

i.e. the envelope consists of the two parallel curves

$$c_{\pm r} = \gamma(t) \pm r\mathbf{N}(t).$$

II. The envelope of the normals: Here the unfolding is

$$F(x,t) = (x - \gamma(t)|\mathbf{T}(t)).$$

The curve $F(x,t_0) = 0$ is the normal to γ at $\gamma(t_0)$. The envelope is the set E_F of those x for which a t exists with

$$(x - \gamma(t)|\mathbf{T}(t)) = 0 \quad \text{and} \quad \kappa(t)(x - \gamma(t)|\mathbf{N}(t)) = 1.$$

i.e. the set of centres of curvature of the curve. In other words, the envelope is the locus of the centres of curvatures i.e. the evolute of the curve.
III. The envelope of the family of circles which have their centres on the curve and pass through the origin. Here the unfolding is given by the function

$$F(x,t) = (\gamma(t) - x|\gamma(t)) - x) - (\gamma(t)|\gamma(t))$$

The envelope is the set of x so that $x = 2(\gamma(t)|\mathbf{N}(t))\mathbf{N}(t)$ for some t i.e. it is the orthotomic. The classical example is the cardioid which is the envelope of the family of those circles with centre on a given circle (passing through the origin) which also pass through the origin.
IV. **Caustics:** The most famous examples of envelopes are the so-called caustics. These are the envelopes of the reflections of rays of light from a point source after being reflected in a given curve. Hence it is the envelope of the normals to the orthotomic of the curve i.e. the evolute of the orthotomic.

Examples of caustics are: The caustic of a circle with the source a point on the circle is the cardioid. With respect to a point not on the circle, it is a limaçon. The caustic of a cardioid with source at its cusp is a nephroid.

5 SURFACES IN SPACE

In our systematic treatment of surfaces, it will be convenient to use the definition involving parametrisation.

Definition: A regular parametrised surface is a C^r-mapping ϕ from an open subset U of the plane into $\mathbf{R}^3$ whose derivative $D\phi$ has rank two at each

point. This means that the partial derivatives ϕ_1 and ϕ_2 are linearly independent. If, further, the parametrisation ϕ is injective, then we call it a **local surface**.

The following are conequences of the definition and various forms of the inverse function theorem:
1) there is a smooth change of coordinates in $\mathbf{R}^3$ so that the surface has locally the trivial form

$$u \mapsto (u_1, u_2, 0)$$

(more precisely, there is a diffeomorphism ψ from a neighbourhood of the range of ϕ onto an open subset of $\mathbf{R}^3$ so that

$$\psi \circ \phi(u) = (u_1, u_2, 0).$$

2) if ϕ is a surface, then every point $u \in U$ has a neighbourhood V in U so that the restriction of ϕ to V is injective (and so a local surface);
3) if $F : V \to \mathbf{R}$ is a C^r-mapping from an open subset V of space into $\mathbf{R}$ which is such that the derivative $(DF)_x$ is non-zero at each point x of V and if $x_0 \in V$ is such that $F(x_0) = 0$, then there is a neighbourhood V_0 of x_0 in V and a local surface $\phi : U \to \mathbf{R}^3$ so that

$$\phi(U) = \{x \in V_0 : F(x) = 0\}.$$

If, in particular, $(D_3F)(x_0) \neq 0$, then ϕ can be chosen of the form

$$u \mapsto (u_1, u_2, f(u))$$

for a smooth function f (i.e. the local surface is the graph of this function).

A useful method of visualising surfaces is by means of the coordinate network which the parmetrisation generates. By this we mean the families

$$t \mapsto \phi(t, u_2) \quad t \mapsto \phi(u_1, t)$$

of curves, whereby the u_1 and u_2 are fixed values of the parameters. As the latter vary, the curves form a two-parameter family which covers the surface—these are the **curvilinear coordinate axes**. Their tangents are proportional to the vectors $\phi_1(u)$ and $\phi_2(u)$. If we have a general curve on the surface through $p = \phi(u)$, then this has (at least locally) the form $\tilde{c} = \phi \circ c$ where c is a curve in U. The derivative of the latter is

$$\dot{\tilde{c}} = \phi_1(u)\dot{c}_1 + \phi_2(u)\dot{c}_2.$$

In other words, the tangent to any curve on the surface through p at this point lies in the plane through p parallel to the one spanned by the vectors $\phi_1(u)$ and $\phi_2(u)$. We call this the **tangent plane** to the surface at p, written $T_p(S)$ or simply T_p.

The Gauß mapping: Since the vectors $\phi_1(u)$ and $\phi_2(u)$ span the tangent plane at p, the unit vector

$$\mathbf{N}(u) = \frac{\phi_1(u) \times \phi_2(u)}{|\phi_1(u) \times \phi_2(u)|}$$

is the unit normal to this plane i.e. it is normal to the surface at p. The function $u \mapsto \mathbf{N}(u)$ is called the **Gauß-mapping** of the surface. The triple $(\phi_1(u), \phi_2(u), \mathbf{N}(u))$ forms a basis for $\mathbf{R}^3$ which need not, however, be orthonormal or even orthogonal.,

If a surface is described as the zero-set of a suitable function F, then the normal at a point x_0 is proportional to the gradient vector of F. For if ϕ is a parametrisation of the surface near x_0, then $F \circ \phi = 0$ and differentiating gives the equation

$$0 = (\operatorname{grad} F | \phi_1) = (\operatorname{grad} F | \phi_2)$$

i.e. the vector $\operatorname{grad} F$ is perpendicular to the tangent plane.

The first fundamental form: We shall be interested in measuring such quantities as the lengths of curves and the areas of sections on surfaces. Of course, we use the standard length resp. scalar product in space. However, owing to the special role played by the coordinate system $(\phi_1(u), \phi_2(u), \mathbf{N}(u))$, we shall employ its matrix representation with respect to this basis and not the canonical basis. In fact, we shall only require the restriction of the bilinear form to the tangent plane, which has as a basis the pair $(\phi_1(u), \phi_2(u))$. The matrix of the scalar product with respect to this basis is

$$A = \begin{bmatrix} E_u & F_u \\ F_u & G_u \end{bmatrix}$$

where

$$E_u = (\phi_1(u)|\phi_1(u)),\ F_u = (\phi_1(u)|\phi_2(u)),\ G_u = (\phi_2(u)|\phi_2(u)).$$

This means that if vectors x and y have representations $x = \lambda_1\phi_1 + \lambda_2\phi_2$ resp. $y = \mu_1\phi_1 + \mu_2\phi_2$, then

$$(x|y) = [l_1\, l_2] A \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}.$$

The quadratic form on the tangent space with this matrix is called the **first fundamental form**. We write $I_u(x, y)$ for the value of the corresponding bilinear form at two tangent vectors x and y. (Of course, this is just the standard scalar product of x and y regarded as vectors in spacel). Hence, in order to calculate the length of a curve $\tilde{c} = \phi \circ c$, we proceed as follows:

$$\dot{\tilde{c}} = (\phi_1 \circ c)\dot{c}_1 + (\phi_2 \circ c)\dot{c}_2$$

and so

$$|\dot{\tilde{c}}|^2 = |(\phi_1 \circ c)\dot{c}_1 + (\phi_2 \circ c)\dot{c}_2|^2 = E_c\dot{c}_1^2 + 2F_c\dot{c}_1\dot{c}_2 + G_c\dot{c}_2^2$$

and so the length is given by the integral

$$I(c) = \int_{t_0}^{t_1} \sqrt{E_{c(t)}\dot{c}_1(t)^2 + 2F_{c(t)}\dot{c}_1(t)\dot{c}_2(t) + G_{c(t)}\dot{c}_2(t)^2}dt$$

In the same way, we can calculate the angle between two curves on a surface as follows: suppose that the curves are $\tilde{\alpha} = \phi \circ \alpha$ and $\tilde{\beta} = \phi \circ \beta$ and that they meet at the point p on the surface. In order to keep the notation simple, we assume that $0 \in U$ and that $p = \phi(0)$, $\alpha(0) = 0 = \beta(0)$. Then if θ is the angle between $\tilde{\alpha}$ and $\tilde{\beta}$ at p, we have

$$\cos\theta = \frac{(\dot{\tilde{\alpha}}(0)|\dot{\tilde{\beta}}(0))}{|\dot{\tilde{\alpha}}(0)||\dot{\tilde{\beta}}(0)|} = \frac{I_0(\dot{\alpha}(0), \dot{\beta}(0))}{\sqrt{I_0(\dot{\alpha}(0), \dot{\alpha}(0))I_0(\dot{\beta}(0), \dot{\beta}(0))}}.$$

Similarly, the area of a local surface, parametrised by the smooth function ϕ defined on the region U is defined by the equation

$$A(M) = \int\int_U H(u)du$$

where $H(u) = |\phi_1(u) \times \phi_2(u)|$.

Note that this definition is independent of the particular parametrisation chosen.

It follows immediately from its definition that the form I is positive definite. In particular, the determinant of its matrix $EG - F^2$ is positive. In fact, by the Lagrange identity

$$(x \times y|z \times u) = (x|z)(y|u) - (y|z)(x|u)$$

we have

$$EG - F^2 = |\phi_1 \times \phi_2|^2 > 0.$$

We shall discuss the geometrical significance of this form later. Of course, the matrix of the first fundamental form depends on the parametrisation used and we discuss briefly the transformation laws satisfied by it.

Suppose then that we have two parametrisations ϕ and $\tilde{\phi}$ which are related by the reprametrisation ψ i.e. $\tilde{\phi} = \phi \circ \psi$ where ψ is a diffeomorphism between the domains of definition of ϕ and $\tilde{\phi}$ with $\det(D\psi) > 0$. In order to make this dependence explicit, we write $E^{\tilde{\phi}}$ resp. E^{ϕ} for the corresponding coefficients. Then, by the chain rule

$$\tilde{\phi}_1 = (\phi_1 \circ \psi)\frac{\partial\psi^1}{\partial u} + (\phi_2 \circ \psi)\frac{\partial\psi^2}{\partial u},$$

$$\tilde{\phi}_2 = (\phi_1 \circ \psi)\frac{\partial\psi^1}{\partial v} + (\phi_2 \circ \psi)\frac{\partial\psi^2}{\partial v}.$$

In particular, $\tilde{\phi}_1(u)$ and $\tilde{\phi}_2(u)$ lie on the tangent plane at p defined by the parametrisation ϕ which shows that the latter is independent of the choice of parametrisation. We have

$$\tilde{\phi}_1 \times \tilde{\phi}_2 = (\phi_1 \times \phi_2) \circ \psi \, \det(D\psi)_{(u,v)}.$$

Since the determinant of the Jacobi matrix of ψ is positive, it follows that $\mathbf{N}^{\tilde{\phi}}(u) = \mathbf{N}^{\phi}(\psi(u))$ i.e. the Gauß-mapping is also independent of the choice of parametrisation. (The same argument shows that if we employ a change of parametrisation with negative determinant, then this reverses the direction of the Gaußian mapping). Now by the definition we have $E^{\tilde{\phi}} = (\tilde{\phi}_1|\tilde{\phi}_1)$ and if we substitute the above expression and simplify we get the equation

$$E^{\tilde{\phi}} = E^{\phi} \circ \psi \Big(\frac{\partial \psi^1}{\partial u}\Big)^2 + 2F^{\phi} \circ \psi \Big(\frac{\partial \psi^1}{\partial u}\frac{\partial \psi^2}{\partial u} + \frac{\partial \psi^2}{\partial u}\partial \psi^1 \partial u\Big) + G^{\phi} \circ \psi \Big(\frac{\partial \psi^2}{\partial u}\Big)^2$$

Similar expressions can be obtained for $F^{\tilde{\phi}}$ and $G^{\tilde{\phi}}$.

We note that in classical notation, these equations take on the form:

$$g^{\tilde{\phi}}_{ik} = \sum_{l,m} g^{\phi}_{lm} \partial \psi^l \partial u_i \partial \psi^m \partial u_k$$

The curvature of a surface: Now consider a point $p = \phi(u)$ on a local surface M with parametrisation $\phi : U \to \mathbf{R}^3$. We wish to investigate geometrical properties of M by considering the curves on the surface obtained by cutting it with suitable planes. For each unit vector x in the tangent plane $T_p(M)$ at p, we consider the curve c_x obtained by intersecting M with the plane spanned by $\mathbf{N}(u)$, the normal vector at p, and x. We denote by κ_x the curvature of this curve at p. In order to deal with this situation analytically, it is convenient to choose a parametrisation of the form

$$\phi : u \mapsto (u_1, u_2, f(u))$$

whereby $p = \phi(0) = 0$ and the partial derivatives of f at the origin are zero. Geometrically, this means that we have chosen coordinates so that the point we are interested in is the origin and the tangent plane there is the horizontal coordinate plane. Suppose now that x is the unit vector $(\cos\theta, \sin\theta)$. Then c_x is the curve

$$t \mapsto (t\cos\theta, t\sin\theta, f(t\cos\theta, t\sin\theta))$$

and a simple calculation shows that

$$2\kappa_{c_x} = \frac{\partial^2 f}{\partial x^2}(0,0)\cos^2\theta + 2\frac{\partial^2 f}{\partial x \partial y}(0,0)\cos\theta\sin\theta + \frac{\partial^2 f}{\partial y^2}(0,0)\sin^2\theta.$$

Proposition 5.1 *Suppose that the curvatures* $\{\kappa_x : x \in S^1\}$ *at a point* p *are not constant. Then there are two values* x_1 *and* x_2 *for which the curvature is a maximum resp. a minimum. Further, these two vectors are perpendicular to each other and if* x *is a third unit vector which makes an angle* θ *with* x_1*, then*

$$\kappa_x = \kappa_{x_1} \cos^2 \theta + \kappa_{x_2} \sin^2 \theta.$$

PROOF. We choose a coordinate system as above. In addition, we can suppose that $\dfrac{\partial^2 f}{\partial u_1 \partial u_2} = 0$ at zero (this follows from elementary linear algebra—we rotate the axes so that the quadratic part of the Taylor expansion of f at zero is in canonical form).
Then if we put $x_1 = (1, 0)$ and $x_2 = (0, 1)$, we get:

$$2\kappa_{x_1} = \frac{\partial^2 f}{\partial u_1^2}(0, 0),$$

$$2\kappa_{x_2} = \frac{\partial^2 f}{\partial u_2^2}(0, 0),$$

$$2\kappa_x = \frac{\partial^2 f}{\partial u_1^2}(0, 0) \cos^2 \theta + \frac{\partial^2 f}{\partial u_2^2}(0, 0) \sin^2 \theta$$

and this implies the result.

■

Using these results, we can give the following definition:

Definition: Let p be a point on a surface M. p is an **umbilical point** if κ_x is constant at p. Otherwise the **principal directions** at p are those vectors x_1 and x_2 in $S^2 \cap T_p(M)$ at which the normal curvature κ_x takes on its minimum (resp. maximum) value κ_1 (resp. κ_2).

We define the following quantities $\kappa = \kappa_1 \kappa_2$ (the **Gaußian curvature**) and $h = \dfrac{(\kappa_1 + \kappa_2)}{2}$ (the **mean curvature**).

We now discuss the fundamental form in the case of a general parametrisation.

The second fundamental form:
If M is a local surface with parametrisation $\phi : U \to \mathbf{R}^3$, we define

$$L^\phi(u) = (\mathbf{N}(u)|\phi_{11}(u)) \quad M^\phi(u) = (\mathbf{N}(u)|\phi_{12}(u)) \quad N^\phi(u) = (\mathbf{N}(u)|\phi_{22}(u)).$$

We shall drop the superscript ϕ when no confusion is possible. The second fundamental form is the bilinear form II_u with L, M, N as coefficients. This means that

$$II_u(x, y) = L\xi_1\eta_1 + M(\xi_1\eta_2 + \xi_2\eta_1) + N\xi_2\eta_2 \tag{1}$$

$$= [\xi_1\, \xi_2] \begin{bmatrix} L & M \\ M & N \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \tag{2}$$

$Then we have the following formulae:$

$\mathrm{L} = 1\frac{}{H(\phi_1|\phi_2\times\phi_{11})=-(\mathbf{N}_1|\phi_1)}Similarly, M = \frac{1}{H}(\phi_1|\phi_2\times\phi_{12}) = -(\mathbf{N}_1|\phi_2) and N = \frac{1}{H}(\phi_1|\phi_2\times\phi_{22}) = -(\mathbf{N}_2|\phi_2). It is sometimes convenient to write \mathrm{L}_{11}, L_{12} = L_{21}, L_{22}$ for L, M, N. Then these formulae take on the form

$$L_{ik} = (\mathbf{N}|\phi_{ik}) = \frac{1}{H}(\phi_1|\phi_2\times\phi_{ik}) = -(\mathbf{N}_i|\phi_k).$$

The L_{ik} satisfy the following transformation laws (where $\tilde{\phi} = \phi\circ\psi$ is a reparametrisation of M):

$$L^{\tilde{\phi}}_{ik} = \sum_{l,m=1}^{2}(L^{\phi}_{lm}\circ\psi)D_i\psi^l D_k\psi^m.$$

This means that (L_{ik}) satisfies the same transformation laws as the coefficients of the first fundamental form i.e. they form a second order covariant tensor on M.

Geometrical background to the second form:
Consider the perpendicular distance $\rho(u)$ from a point $\phi(u)$ on M to the tangent plane at a given point p where we choose ϕ so that $p = \phi(0,0)$. Then

$$\rho(u) = (\phi(u) - \phi(0)|\mathbf{N}(0))$$

and if we consider the Taylor expansion

$$\phi(u) = \phi(0) + \phi_1(0)u_1 + \phi_2(0)u_2$$

$$+\frac{1}{2}\big[\phi_{11}(0)u_1^2 + 2\phi_{12}(0)u_1u_2 + \phi_{22}(0)u_2^2\big] + \dots$$

we have

$$\rho(u) = \frac{1}{2}\Big[\big(\phi_{11}(0)|\mathbf{N}(0)\big)u_1^2 + 2\big(\phi_{12}(0)|\mathbf{N}(0)\big)u_1u_2 + \big(\phi_{22}(0)|\mathbf{N}(0)\big)u_2^2\Big] + \dots$$

$$= \frac{1}{2}[L_0u_1^2 + 2M_0u_1u_2 + N_0u_2^2] + \dots$$

i.e. the second fundamental form is the best quadratic approximation to the surface at the given point (after rotating to bring its tangent plane into the horizontal position).

Using this, we can classify a point $p = \phi(u)$ as

1) **elliptic** if II_u is positive or negative definite (i.e. $L(u)N(u) - M^2(u) > 0$);

2) **hyperbolic** if II_u is indefinite (i.e. $L(u)N(u) - M^2(u) < 0$);

3) **parabolic** if $L(u)N(u) - M^2(u) = 0$ but at least one of the coefficients is non-zero;

4) **planar** if all of the coefficients vanish.

Coordinate free definition of the second form: Consider the surface parametrised by the Gaußian mapping $\mathbf{N}$ which is a smooth mapping from U into the unit sphere (of course this surface can have singularities). The derivative $(D\mathbf{N})_u$ of this mapping at u is a linear mapping from the subset U of the plane into three-space. Since the length of the normal vector is always one, then, by differentiating the equation $(\mathbf{N}|\mathbf{N}) = 1$, we obtain the relationships $(\mathbf{N}|\mathbf{N}_1) = (\mathbf{N}|\mathbf{N}_2) = 0$ i.e. the vectors $\mathbf{N}_1$ and $\mathbf{N}_2$ are perpendicular to $\mathbf{N}$. Hence they are in the tangent plane $T_p(M)$. We can thus define a bilinear form on the plane as follows:

$$(x, y) \mapsto -((D\mathbf{N})x|(D\phi)y).$$

This is symmetric. For, differentiating the equation $(\mathbf{N}|\phi_1) = 0$, we obtain the equations

$$-(\mathbf{N}_2|\phi_1) = (\mathbf{N}|\phi_{12}) = -(\mathbf{N}_1|\phi_2).$$

With respect to the standard basis (e_1, e_2) for the plane, the bilinear form has coefficients (L, M, N) i.e. it is precisely the form II_u.

The shape operator: The shape operator S_u at the point u is the operator

$$-(D\mathbf{N})_u \circ (D\phi)_u^{-1}$$

on the tangent plane. This operator defines a bilinear form $(x, y) \mapsto (S_u x|y)$ thereon. Note that if x, y are in $\mathbf{R}^2$, then

$$II_u(x, y) = \big(S_u(D\phi)_u x|(D\phi)_u y\big)$$

i.e. the bilinear form defined on $T_p(M)$ induces the second fundamental from on $\mathbf{R}^2$ via the mapping $(D\phi)_u$.

Hence the second fundamental form has appeared in the following essentially equivalent disguises:

a) as a bilinear form on the plane, defined as follows:

$$(x, y) \mapsto \big(S_u(D\phi)_u x, (D\phi)_u y\big)$$

b) as a symmetric linear mapping S_u on the tangent plane

c) as a bilinear form

$$(x, y) \mapsto \big(S_u x|y\big)$$

on the tangent plane.

The matrix of II_u as a bilinear form on the plane with respect to the standard basis is $[l_{ij}]$. Hence this is also the basis of the second form as a bilinear form on the tangent space with respect to the basis (ϕ_1, ϕ_2). It follows from elementary linear algebra that the operator S has the following matrix with respect to (ϕ_1, ϕ_2).

$$\begin{bmatrix} E & F \\ F & G \end{bmatrix}^{-1} \begin{bmatrix} L & M \\ M & N \end{bmatrix} = \frac{1}{H^2} \begin{bmatrix} GL - FM & GM - FN \\ -FL + EM & -FM + EN \end{bmatrix}$$

We can now give a linear-algebraic proof of the results on the principal curvature from the beginning of the chapter. Supose that the shape operator has eigenvalues κ_1, κ_2 with corresponding eigenvectors x_1, x_2. Then the latter are perpendicular since S is symmetric. Further if $x = x_1 \cdot \cos\theta + x_2 \cdot \sin\theta$ is a unit vector on the tangent plane, then

$$\kappa_x = \big(S_p x | x\big)$$

In particular, it follows that the principal curvatures κ_1 and κ_2 are just the eigenvalues of S and so can be calculated as the roots of the characteristic polynomial of the matrix

$$\frac{1}{H^2}\begin{bmatrix} GL - FM & GM - FN \\ -FL + EM & -FM + EN \end{bmatrix}$$

The sum of these roots is $\dfrac{GL - F + EN - FM}{H^2}$ and their product is

$$\frac{(GL - FM)(-FM + EN) - (FL - EM)(-GM + FN)}{H^4}.$$

Simplifying, we get the formulae

$$\kappa = \frac{LN - M^2}{H^2} = \frac{LN - M^2}{EG - F^2}$$

$$h = \frac{1}{2}\frac{LG + EN - 2MF}{H^2} = \frac{1}{2}\frac{LG + EN - 2FM}{EG - F^2}$$

$$\kappa_1 = h - \sqrt{(h^2 - \kappa)},$$

$$\kappa_2 = h + \sqrt{(h^2 - \kappa)}.$$

The Theorema egregium: The Gaussian curvature κ is given by the formula

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

and so depends apparently on (the coefficients of) both fundamental forms. In fact, as the following formula shows, it only depends on those of the first form (together with their derivatives), a fact which has far reaching geometrical consequences:

Proposition 5.2 *We have the following formula for the curvature κ of a surface:*

$$\begin{aligned} 4(EG - F^2)^2\kappa = {} & E\big(E_2G_2 - 2F_1G_2 + (G_1)^2\big) + && (3)\\ & + F(E_1G_1 - E_2G_1 - 2E_2F_2 + 4F_1F_2 - 2F_1G_1) + && (4)\\ & + G(E_1G_1 - 2E_1F_2 + E_2^2) - 2(EG - F^2)(E_{22} - 2F_{12} + G_{11}). && (5) \end{aligned}$$

PROOF. We have $\kappa = \dfrac{(LN - M^2)}{(EG - F^2)}$ where

$$L = (\phi_{11}|N) = \frac{(\phi_{11}|\phi_1 \times \phi_2)}{(EG - F^2)^{\frac{1}{2}}} \tag{6}$$

$$M = \frac{(\phi_{12}|\phi_1 \times \varphi y_2)}{(EG - F^2)^{\frac{1}{2}}} \tag{7}$$

$$N = \frac{(\phi_{22}|\phi_1 \times \phi_2)}{(EG - F^2)^{\frac{1}{2}}}. \tag{8}$$

$$\tag{9}$$

Hence

$$\kappa(EG - F^2)^2 = (\phi_{11}|\phi_1 \times \phi_2)(\phi_{22}|\phi_1 \times \phi_2) - (\phi_{12}|\phi_1 \times \phi_2)^2$$

$$= \det \begin{bmatrix} \phi_{11} \\ \phi_1 \\ \phi_2 \end{bmatrix} \det \begin{bmatrix} \phi_{22} \\ \phi_2 \\ \phi_2 \end{bmatrix} - \det \begin{bmatrix} \phi_{12} \\ \phi_1 \\ \phi_2 \end{bmatrix} \det \begin{bmatrix} \phi_{12} \\ \phi_1 \\ \phi_2 \end{bmatrix}$$

(where we are taking the determinant of the matrices whose rows consist of the vectors $\phi_{11}, \phi_1, \phi_2$, etc.)

$$= \det \begin{bmatrix} \phi_{11} \\ \phi_1 \\ \phi_2 \end{bmatrix} \det(\phi_{22}^t, \phi_1^t, \phi_2^t) - \det \begin{bmatrix} \phi_{12} \\ \phi_1 \\ \phi_2 \end{bmatrix} \det(\phi_{12}^t, \phi_1^t, \phi_2^t)$$

(where we have transposed two of the matrices so that the column vectors are now $\phi_{22}^t, \phi_1^t, \phi_2^t$ etc. i.e. the row vectors $\phi_{22}, \phi_1, \phi_2$ written as columns)

$$= \det \begin{bmatrix} (\phi_{11}|\phi_{22}) & (\phi_{11}|\phi_1) & (\phi_{11}|\phi_2) \\ (\phi_1|\phi_{22}) & E & F \\ (\phi_2|\phi_{22}) & F & G \end{bmatrix}$$

$$- \det \begin{bmatrix} (\phi_{12}|\phi_{12}) & (\phi_{12}|\phi_1) & (\phi_{12}|\phi_2) \\ (\phi_{12}|\phi_1) & E & F \\ (\phi_{12}|\phi_2) & F & G \end{bmatrix}$$

$$= \big((\phi_{11}|\phi_{22}) - (\phi_{12}|\phi_{12})\big) \det \begin{bmatrix} E & F \\ F & G \end{bmatrix}$$

$$+ \det \begin{bmatrix} 0 & (\phi_{11}|\phi_1) & (\phi_{11}|\phi_2) \\ (\phi_1|\phi_{22}) & E & F \\ (\phi_2|\phi_{22}) & F & G \end{bmatrix}$$

$$- \det \begin{bmatrix} 0 & (\phi_{12}|\phi_1) & (\phi_{12}|\phi_2) \\ (\phi_{12}|\phi_1) & E & F \\ (\phi_{12}|\phi_2) & F & G \end{bmatrix}$$

Now $E = (\phi_1|\phi_1)$ and so $E_1 = 2(\phi_{11}|\phi_1)$ i.e. $(\phi_{11}|\phi_1) = \frac{1}{2}E_1$.

Similarly,

$$(\phi_{12}|\phi_1) = \frac{1}{2}E_2, \quad (\phi_{22}|\phi_2) = \frac{1}{2}G_2. \tag{10}$$

$$(\phi_{12}|\phi_2) = \frac{1}{2}G_1, \quad (\phi_{11}|\phi_2) = F_1 - \frac{1}{2}E_2. \tag{11}$$

$$(\phi_{22}|\phi_1) = F_2 - \frac{1}{2}G_1. \tag{12}$$

$$\tag{13}$$

Also, differentiating the formulae for E, F, G we get:

$$E_2 = 2(\phi_{12}|\phi_1), E_{22} = 2(\phi_{122}|\phi_1) + 2(\phi_{12}|\phi_{12}), \tag{14}$$

$$F_1 = (\phi_{11}|\phi_2) + (\phi_1|\phi_{12}), \tag{15}$$

$$F_{12} = (\phi_{112}|\phi_2) + (\phi_{11}|\phi_{22}) + (\phi_{12}|\phi_{12}) + (\phi_1|\phi_{122}), \tag{16}$$

$$G_1 = 2(\phi_{12}|\phi_2), G_{11} = 2(\phi_{112}|\phi_2) + 2(\phi_{12}|\phi_{12}) \tag{17}$$

$$\tag{18}$$

and so

$$(E_{22} - 2F_{12} + G_{11}) = -2\big((\phi_{11}|\phi_{22}) - (\phi_{12}|\phi_{12})\big).$$

The proof is now completed by a routine calculation which we shall omit since the essential point, namely that κ can be expressed in terms of E, F, G and their partial derivatives, is now apparent.

■

Principal directions: Let M be a local surface, $p = \phi(u) \in M$. We suppose that p is not an umbilical point i.e. that $\kappa_1(u) < \kappa_2(u)$ where these are the principal curvatures. If the surface has no umbilical points, then the above functions κ_1 and κ_2 are smooth functions on the parametrising space U. A curve $\tilde{c} = \phi \circ c$ on M is called a **line of curvature** if $\kappa_{\tilde{c}} = \kappa_1 \circ c$ or $\kappa_{\tilde{c}} = \kappa_2 \circ c$ i.e. if the curvature at each point on c is one of the principal curvatures. For example, the coordinate curves are lines of curvature whenever $F = M = 0$ i.e. the bases (ϕ_1, ϕ_2) are orthogonal at each point and the matrix of S_u with respect to these bases is diagonal.

If M fails to have umbilical points, then one can always choose (locally) a parametrisation so that the coordinate curves are lines of curvature. This is a consequence of the following general result whose proof uses existence theorems for partial differential equations.

Proposition 5.3 *Let $\phi : U \to M$ be a parametrised surface and suppose that $X_1 : U \to T(M)$, $X_2 : U \to T(M)$ are smooth functions with the property that*

$X_1(u)$ and $X_2(u)$ are in $T_{\phi(u)}$ and are linearly independent there for each u. (Such functions are called **tangential vector fields** *– they will be studied in more detail in the final chapter). Then for each $u_0 \in U$ there is a neighbourhood V of u_0 in U and a reparametrisation $\tilde{\phi}$ of $\phi|_V$ so that $\tilde{\phi}_1$ is parallel to X_1 and $\tilde{\phi}_2$ is parallel to X_2 on this neighbourhood.*

As noted above, the result on the existence of a parametrisation which has the coordinate curves as lines of curvature holds for surfaces without umbilical points. In general, if a surface does not contain parts which have the form of a portion of a sphere or a plane, then umbilical points occur in isolation. More precisely:

Proposition 5.4 *Let $\phi : U \to M$ be a local surface for which every point is umbilical. Then either*

a) M is planar (i.e. the function $\mathbf{N}$ is constant)

or

b) M is spherical (i.e. there is a point x_0 so that $|\phi - x_0|$ is constant).

PROOF. First we note that $p = \phi(u)$ is umbilical if and only if $\{(S_u x|x) : x \in S^2 \cap T_p(M)\}$ is constant i.e. every vector in the tangent space is an eigenvector of S_u. By an elementary result of linear algebra, this can only happen if the shape operator is a constant times the identity i.e. if $S_u = \kappa(u)\mathrm{Id}$. This means that $(D\mathbf{N})_u = -\kappa(u)(D\phi)_u$ and so $\mathbf{N}_1 + \kappa\phi_1 = 0$ and $\mathbf{N}_2 + \kappa\phi_2 = 0$. Hence $\mathbf{N}_{12} = -\kappa_2\phi_1 - \kappa\phi_{12} = -\kappa_1\phi_2 - \kappa\phi_{12}$. Thus $\kappa_2\phi_1 = \kappa_1\phi_2$ and so $\kappa_1 = \kappa_2 = 0$ (since ϕ_1 and ϕ_2 are linearly independent) i.e. κ is constant.
Case 1: κ is zero. Then $D\mathbf{N} = 0$ i.e. $\mathbf{N}$ is constant;
Case 2: κ is non-zero. Then

$$(\frac{\mathbf{N}}{\kappa} + \phi)_1 = (\frac{\mathbf{N}}{\kappa} + \phi)_2 = 0$$

i.e. $\frac{\mathbf{N}}{\kappa} + \phi = x_0$ for some $x_0 \in \mathbf{R}_3$ and then $|\phi - x_0| = |\frac{\mathbf{N}}{\kappa}| = \frac{1}{\kappa}$.

■

The geodetic curvature: Consider a curve $\tilde{c} = \phi \circ c$ on M, with $|\dot{\tilde{c}}| = 1$. We have $\ddot{\tilde{c}}(t) = \kappa_{\tilde{c}}(t)\mathbf{N}_{\tilde{c}}(t)$. Now we split the "acceleration" $\ddot{\tilde{c}}$ into its component along $\mathbf{N}$ and its component on $T_p(M)$ i.e we write

$$\ddot{\tilde{c}}(t) = l_1\mathbf{N}\big(c(t)\big) + l_2 u(t)$$

where $l_1, l_2 \in \mathbf{R}$ and $u(t)$ is a unit vector in $T_p(M)$. We calculate $l_1, l_2, u(t)$ as follows: firstly

$$l_1 = \big(\ddot{\tilde{c}}(t)|\mathbf{N}(c(t))\big) = \kappa_x \quad \text{the normal curvature in the direction} \quad x = \ddot{\tilde{c}}(t).$$

Also $\dot{\tilde{c}} \perp \ddot{\tilde{c}}$ and so the component of $\mathbf{N}_c$ on $T_p(M)$ is perpendicular to $\mathbf{N}$ and to $\dot{\tilde{c}}$ i.e. $\mathbf{N} \times \dot{\tilde{c}}$ is a suitable candidate for u. Then we have $l_2 = \big(\ddot{\tilde{c}} | (\mathbf{N} \circ c) \times \dot{\tilde{c}}\big)$ and we denote this latter quantity by κ_g—the **geodetic curvature** of $\tilde{c}$ at t.

We have the formulae:

$$\kappa_{\tilde{c}} \mathbf{N}_{\tilde{c}} = \kappa_x \mathbf{N} \circ c + \kappa_g \big((\mathbf{N} \circ c) \times \dot{\tilde{c}} \big) \tag{19}$$

$$\kappa_{\tilde{c}}^2 = \kappa_x^2 + \kappa_g^2. \tag{20}$$

Note that, with respect to an arbitrary parametrisation (i.e. without the assumption $|\dot{\tilde{c}}| = 1$), we have the formula:

$$\kappa_g = (\ddot{\tilde{c}} | \dot{\tilde{c}} \times \mathbf{N} \circ c) |\dot{\tilde{c}}|^3$$

A curve $\tilde{c}$ is a **geodetic** if $\kappa_g = 0$.
For example, a straight line on a surface is always a geodetic (for then $\kappa_{\tilde{c}}^2 = 0$ and so $\kappa_g^2 = 0$). The geodetics on the sphere are curves which lie on great circles. Normally one thinks of a geodetic as that curve between two points on a surface which has the shortest length. In fact, such curves are geodetics in the above sense. The converse is not true as can be seen from the example of two segments of a great circle between points on the sphere.

The precise relationship between our local definition of geodetics and the global definition can be clarified within the framework of the calculus of variations.

Minimal surfaces: A local surface M is **minimal** if its mean curvature $h = 0$. In the spirit of the remarks on geodetics, we note that this definition is related to the fact that M is the surface with the smallest area among all surfaces with the same "boundary". A suitable model is a soap film supported on a frame. Once again, this statement can be made precise within the framework of the calculus of variations. We characterise here those surfaces of revolution

$$\phi \colon (u_1, u_2) \mapsto \big(f(u_1) \cos u_2, f(u_1) \sin u_2, g(u_1) \big)$$

which are minimal. First we assume, as we may, that $(f')^2 + (g')^2 = 1$ (this just means that the generating curve $t \mapsto \big(0, f(t), g(t)\big)$ is parametrised by arc length). Then a routine calculation shows:

$$E = 1, F = 0, G = f^2 \tag{21}$$

$$L = f'g'' - f''g', M = 0, N = fg' \tag{22}$$

Hence

$$\kappa = g'(f'g'' - f''g')f \quad h = g' + f(f'g'' - f''g')2f \tag{23}$$

$$\kappa_1 = g'f, \quad \kappa_2 = f'g'' - f''g'. \tag{24}$$

Hence the surface of revolution is minimal iff $\frac{g'}{f} + (f'g'' - f''g') = 0$. A rather messy calculation shows that this is the case if

$$f(t) = a\cosh\left(\frac{g(t) - b}{a}\right)$$

for suitable constants a, b. Then the surface is a **catenoid** i.e. the surface of revolution generated by a catenary.

Ruled surfaces: Let γ, c be curves in $\mathbf{R}^3$ where, for convenience, we assume that γ is parametrised by arc length and $|c| = 1$. We call the surface

$$\phi\colon u \mapsto \gamma(u_1) + u_2 c(u_1)$$

a **ruled surface**.
The line $t \mapsto \gamma(u_1) + tc(u_1)$ is called **the generator through** $\gamma(u)$. Note that we have:

$$\phi_1\colon (u_1, u_2) \mapsto \dot{\gamma}(u_1) + u_2\dot{c}(u_1) \tag{25}$$
$$\phi_2\colon (u_1, u_2) \mapsto c(u_1) \tag{26}$$
$$\phi_1 \times \phi_2\colon (u_1, u_2) \mapsto \big(\dot{\gamma}(u_1) + u_2\dot{c}(u_1)\big) \times c(u_1) \tag{27}$$

and so the surface is regular at those points where the last expression is non-zero.

Example – Tangent surfaces: The tangent surface to γ is the surface

$$\phi\colon u \mapsto \gamma(u_1) + u_2\mathbf{T}_\gamma(u_1)$$

Then

$$\phi_1 \times \phi_2\colon u \mapsto -u_2\kappa_\gamma(u_1)\mathbf{B}_\gamma(u_1)$$

and so the points of γ are singularities of ϕ.

Similarly, we can define the **normal surface**

$$u \mapsto \gamma(u_1) + u_2\mathbf{N}_\gamma(u_1)$$

and the **binormal surface**

$$u \mapsto \gamma(u_1) + u_2\mathbf{B}_\gamma(u_1)$$

We can easily calculate the first fundamental form of a tangent surface

$$E = 1 + \kappa_\gamma^2(u_1)u_2^2 \tag{28}$$
$$F = 1 \tag{29}$$
$$G = 1. \tag{30}$$

In particular, it is independent of the torsion of γ. Now we know that there exists a plane curve γ_1 with $\kappa_\gamma = \kappa_{\gamma_1}$. Then the tangent surfaces of γ_1 and γ have tha same metric form and so are isometric. Hence we have the result:
The tangent surface of a curve is isometric to a plane surface (since the tangent surface of γ_1 is obviously planar).

The Gaußian curvature as a limit of quotients of areas: The Gaußian curvature can be given an attractive geometric interpretation as follows: Consider the surface $\phi\colon U \to \mathbf{R}^3$ and the associated normal surface $\mathbf{N}\colon U \to \mathbf{R}^3$.

If $u_0 \in U, \epsilon > 0, B_\epsilon(u_0)$ is the ball $\{u \in U\colon |u - u_0| < \epsilon\}$. Then we claim

$$\big|\kappa(u_0)\big| = \lim_{\epsilon \to 0^+} \int\int_{B_\epsilon(u_0)} |\mathbf{N}_1(u) \times \mathbf{N}_2(u)| du \int\int_{B_\epsilon(u_0)} |\phi_1(u) \times \phi_2(u)| du$$

i.e. it is the limit of the quotient of the area traced out by ϕ and $\mathbf{N}$ over $B_\epsilon(u_0)$.

To prove this, we note that, by the mean value theorem, the limit is just

$$\big|\mathbf{N}_1(u_0) \times \mathbf{N}_2(u_0)\big|\big|\phi_1(u_0) \times \phi_2(u_0)\big|$$

Now we have seen that

$$-\mathbf{N}_1 = GL - FMH^2\phi_1 + GM - FNH^2\phi_2 \tag{31}$$
$$-\mathbf{N}_2 = -FL + EMH^2\phi_1 + -FM + ENH^2\phi_2 \tag{32}$$

and so

$$\mathbf{N}_1 \times \mathbf{N}_2 = (GL - FM)(-FM + EN) - (-FL + EM)(GM - FN)H^4(\phi_1 \times \phi_2)$$

and simplifying we get

$$|\mathbf{N}_1 \times \mathbf{N}_2||\phi_1 \times \phi_2| = |LN - M^2|H^2 = |\kappa|$$

The Christoffel symbols: We now investigate the higher derivatives of the parametrisation of a surface. The second derivatives ϕ_{ik} of ϕ have at each point on the surface a representation in terms of the basis $(\phi_1, \phi_2, \mathbf{N})$. The component in the direction $\mathbf{N}$ is described by the coefficients of the second fundamental form. For the other components, we can introduce symbols Γ^l_{ik} which are defined by the equations

$$\phi_{ik} = \sum_l \Gamma^l_{ik}\phi_l + \ell_{ik}\mathbf{N}$$

In the following we use the notation:

$$g_{ik} = \big(\phi_i|\phi_k\big), \ell_{ik} = \big(\mathbf{N}|\phi_{ik}\big) \tag{33}$$
$$G = [g_{ik}] \tag{34}$$
$$G^{-1} = [g^{ik}] \tag{35}$$

i.e.

$$\begin{bmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{bmatrix} = 1 \det G \begin{bmatrix} {}^{22} & \\ -g_{12} & \\ -g_{12} & \\ g_{11} & \end{bmatrix}$$

If we take the scalar product of the defining equations with ϕ_m we get the equation

$$\big(\phi_{ik}|\phi_m\big) = \sum_l \Gamma^l_{ik} g_{lm}$$

i.e.

$$\Gamma^l_{ik} = \sum_m \big(\phi_{ik}|\phi_m\big) g^{lm}$$

The Γ^l_{ik} are called **the Christoffel symbols of the first kind**.

Differentiating the equation

$$g_{im} = \big(\phi_i|\phi_m\big)$$

we get the formula

$$\partial g_{im} \partial u_k = \big(\phi_{ik}|\phi_m\big) + \big(\phi_i|\phi_{mk}\big)$$

If we substitute this in the right hand side of the following, we see that

$$\big(\phi_{ik}|\phi_m\big) = 12\left(\partial g_{im}\partial u_k + \partial g_{mk}\partial u_i - \partial g_{ki}\partial u_m\right)$$

We introduce the symbol Γ_{imk} for the left-hand side of this equation (these are called the **Christoffel symbols of the second kind**). Then

$$\Gamma^l_{ik} = \sum_m g^{lm}\Gamma_{ikm}.$$

If we differentiate the equation

$$\phi_{ik} = \sum_l \Gamma^l_{ik}\phi_l + \ell_{ik}\mathbf{N}$$

we get

$$\phi_{ikj} = \sum_n \left[\partial\Gamma^n_{ik}\partial u_j + \sum_l \Gamma^l_{ik}\Gamma^n_{lj} - \ell_{ik}\ell^n_j\right]\phi_n + \left[\partial\ell_{ik}\partial u_j + \sum_l \Gamma^l_{ik}\ell_{lj}\right]\mathbf{N}.$$

Similarly we introduce quantities ℓ^j_i by the equation

$$\mathbf{N}_i = -\sum_j \ell^j_i \phi_j.$$

(Note that the ℓ_i^j are obtained from the ℓ_{ij} where $[\ell_{ij}]$ is the matrix of the second fundamental form by the operation of raising an index.)

Noting the fact that $\phi_{ikj} = \phi_{ijk}$ etc. we get

$$\partial\Gamma^n_{ik}\partial u_j - \partial\Gamma^n_{ij}\partial u_k + \sum_l \Gamma^l_{ik}\Gamma^n_{lj} - \sum_l \Gamma^l_{ij}\Gamma^n_{lk} = \ell_{ik}\ell^n_j - \ell_{ij}\ell^n_k$$

$$\partial\ell_{ik}\partial u_j - \partial\ell_{ij}\partial u_k + \sum_l \Gamma^l_{ik}\ell_{lj} - \sum_l \Gamma^l_{ij}\ell_{lk} = 0.$$

The information contatined in the above equations reduces to the two equations

$$\partial\ell_{12}\partial u_1 - \partial\ell_{11}\partial u_k + \sum_l \Gamma^1_{12}\ell_{l1} - \sum_l \Gamma^1_{11}\ell_{12} = 0$$

$$\partial\ell_{22}\partial u_1 - \partial\ell_{21}\partial u_2 + \sum_l \Gamma^l_{22}\ell_{l1} - \sum_l \Gamma^l_{2l}\ell_{12}$$

(the **Codazzi-Mainardi equations**).

An important consequence of these equations is the fact that the first and second fundamental forms are not independent of each other—in other words, given two forms I_u and II_u (in dependence on u) where the first is positive definite, it need not be true (even locally) that there exists a surface with I and II as fundamental forms. A necesary condition for this is that they satisfy the Codazzi-Mainardi equations.

We denote the left hand side of the above equation i.e.

$$\partial\Gamma^n_{ik}\partial u_j - \partial\Gamma^n_{ij}\partial u_k + \sum_l \Gamma^l_{ik}\Gamma^n_{lj} - \sum_l \Gamma^l_{ij}\Gamma^n_{lk}$$

by R^n_{ijk}. Then, trivially $R^n_{ijk} = -R^n_{ikj}$.
Further we define

$$R_{mijk} = \sum_n g_{mn}R^n_{ijk} (= \ell_{ik}\ell_{jm} - \ell_{ij}\ell_{km})$$

Then note that

$$R_{1212} = \ell_{22}\ell_{11} - (\ell_{21})^2 = \det[\ell_{ij}].$$

Hence we have the following formula for the Gaußian curvature

$$\kappa = R_{1212}g$$

(this is a restatement of the Theorema Egregium).

We derive the equations of a geodetic as follows:
A curve $\dot{\tilde{c}} = \phi \circ c$ on M is a geodetic if $\kappa_g = 0$. This means that $\ddot{\tilde{c}}(t) \| \mathbf{N}_{c(t)}$ i.e.

$$\big(\ddot{\tilde{c}}(t) | \phi_i\big) = 0 \quad \text{for} \quad i = 1, 2.$$

Now

$$\dot{\tilde{c}}(t) = \phi_1\big(c(t)\big)\dot{c}_1(t) + \phi_2\big(c(t)\big)\dot{c}_2(t) \tag{36}$$
$$\ddot{\tilde{c}}(t) = \phi_{11}\big(c(t)\big)\dot{c}_1(t)^2 + 2\phi_{12}\big(c(t)\big)\dot{c}_1(t)\dot{c}_2(t) + \tag{37}$$
$$+ \phi_{22}\big(c(t)\big)\dot{c}_2(t)^2 + \phi_1\big(c(t)\big)\ddot{c}_1(t) + \phi_2\big(c(t)\big)\ddot{c}_2(t) \tag{38}$$
$$= \sum_{i,j} \phi_{ij}\big(c(t)\big)\dot{c}_i(t)\dot{c}_j(t) + \sum_i \phi_i\big(c(t)\big)\ddot{c}_i(t). \tag{39}$$

Hence the equations take on the form

$$\sum_{i,j}\Big(\phi_{ij}\big(c(t)\big)\phi_k\big(c(t)\big)\dot{c}_i(t)\dot{c}_j(t)\Big) + \sum_i\Big(\phi_i\big(c(t)\big)|\phi_k\big(c(t)\big)\ddot{c}_i(t)\Big) = 0$$

i.e.

$$E\ddot{c}_1 + F\ddot{c}_2 + \Gamma_{111}\dot{c}_1^2 + 2\Gamma_{112}\dot{c}_1\dot{c}_2 + \Gamma_{122}\dot{c}_2^2 = 0 \tag{40}$$
$$F\ddot{c}_1 + G\ddot{c}_2 + \Gamma_{211}\dot{c}_1^2 + 2\Gamma_{212}\dot{c}_1\dot{c}_2 + \Gamma_{222}\dot{c}_2^2 = 0 \tag{41}$$

or, solving for $\ddot{c}_1$ and $\ddot{c}_2$,

$$\ddot{c}_1 + \Gamma^1_{11}\dot{c}_1^2 + 2\Gamma^1_{12}\dot{c}_1\dot{c}_2 + \Gamma^1_{22}\dot{c}_2^2 = 0 \tag{42}$$
$$\ddot{c}_2 + \Gamma^2_{11}\dot{c}_1^2 + 2\Gamma^2_{12}\dot{c}_1\dot{c}_2 + \Gamma^2_{22}\dot{c}_2^2 = 0 \tag{43}$$

Example: consider the surface of revolution

$$\phi(u, v) = (u\cos v, u\sin v, f(u)).$$

Then we get the equation

$$\frac{d}{dt}\left(c_1^2\left(\frac{dc_2}{dt}\right)0^2\right) = 0.$$

In particular, the geodetics on a cylinder are helices and on a cone they are concho-spirals i.e. on the surface

$$(u, v) \mapsto (u\cos v, u\sin v, u)$$

they are induced by the curves

$$t \mapsto \frac{1}{a}\sec(\beta + t\sin\alpha), t).$$

6 DIFFERENTIABLE MANIFOLDS

Definition: An (n-dimensional) **topological manifold** is a connected metric space (X, τ) with the property that each point x has an open neighbourhood U which is homeomorphic to (an open subset of) $\mathbf{R}^n$. A homeomorphism ϕ from U onto an open subset of $\mathbf{R}^n$ is called a **chart** for X at x. A family $\{(U_\alpha, \phi_\alpha)\}$ of charts so that the $\{U_\alpha\}$ cover X is called an **atlas**.

Examples:

I. $\mathbf{R}$ and S^1 are examples of one-dimensional manifolds. (In fact, these are the **only** examples of connected, one-dimensional manifolds).

II. If M_1 and M_2 are n_1- resp. n_2-dimensional manifolds, then their product $M_1 \times M_2$ is a manifold of dimension $(n_1 + n_2)$. For example $S^1 \times S^1$ is the standard torus. An n-fold product of copies of S^1 is the **n-torus.**

III. The **Bretzel:** see the figure. More generally, we have the n-holed torus further examples of two-dimensional manifolds are displayed in the figures.

An **n-dimensional differentiable manifold** (more precisely, a C^r-manifold where $1 \leq r \leq \infty$) is a topological manifold M, together with an atlas $\{(U_\alpha, \phi_\alpha)\}$ so that whenever the intersection $U_\alpha \cap U_\beta$ of two charts is non-empty, then the mapping $\phi_\beta \circ \phi_\alpha^{-1}$ from $\phi_\alpha(U_\alpha \cap U_\beta)$ onto $\phi_\beta(U_\alpha \cap U_\beta)$ is smooth (more precisely, C^r.) It follows then by symmetry that the above mapping, which describes the change of coordinates from one chart to the other, is a diffeomorphism.

Examples:

I. If U is an open subset of $\mathbf{R}^n$, then it is in a trivial way a manifold. Manifolds of this type are called **local manifolds.**

II. We can provide $\mathbf{R}$ with a second structure by using the mapping $t \mapsto t^3$ as a chart. The corresponding structure on $\mathbf{R}$ does not coincide with the natural one.

III. The n-sphere S^n: The usual way for providing this set with an atlas is by means of stereographic projection i.e. we define

$$U_N = S^n \backslash \{N\} \quad U_S = S^n \backslash \{S\}$$

where $N = (1, 0, \ldots, 0)$ is the north pole and $S = (-1, 0, \ldots, 0)$ is the south pole. The corresponding coordinate functions are

$$\phi_N : (\xi_0, \ldots, \xi_n) \mapsto \frac{1}{1 - \xi_0}(\xi_1, \ldots, \xi_n)$$

and

$$\phi_S : (\xi_0, \ldots, \xi_n) \mapsto \frac{1}{1 + \xi_0}(\xi_1, \ldots, \xi_n).$$

The transformation function $\phi_S \circ \phi_N^{-1}$ is, in this case, the function $y \mapsto \frac{y}{|y|^2}$.

IV. The n-dimensional **projective space** P^n is defined to be the quotient of the punctured space $\mathbf{R}_{n+1} \backslash \{0\}$ under the equivalence relation

$$x = (\xi_0, \ldots, \xi_n) \sim y = (\eta_0, \ldots, \eta_n) \quad \text{if and only if there is a} \quad c \neq 0 \quad \text{so that} \quad x = cy.$$

P^n is provided with the natural quotient topology. We define an atlas on this space as follows. We denote by V_i the set

$$\mathbf{R}^{n+1}\backslash\{x : \xi_i = 0\}$$

and by U_i the image of V_i under the quotient mapping π. We then define a mapping ϕ_i from U_i into $\mathbf{R}^n$ as follows:

$$\phi_i(\pi(x)) = \frac{(-1)^i}{\xi_i}(\xi_0, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_n).$$

The family of all the (U_i, ϕ_i) is an atlas for the differentiable structure on P^n.
VI. Submanifolds: A subset M_1 of a manifold M is a **submanifold** if for each $x_0 \in U$ there is a chart (U, ϕ) so that $\phi(M_1) = \{0\} \times V$ for some open subset V of $\mathbf{R}^{n-r}$. In this situation, M_1 has a natural manifold structure (we take as an atlas the sets of the form $\tilde{U} = U \cap M_1$ with U as above. The corresponding mapping is ϕ restricted to $\tilde{U}$).

Having defined differentiable manifolds, we now define the appropriate concept of differentiable mappings between them.

Definition: Let M and N be manifolds. A mapping f from M into N is **differentiable** (more precisely, C^r) if for every $x \in X$ there are charts (U, ϕ_U) and (V, ϕ_V) of M at x resp. of N at $y = f(x)$, so that $f(U) \subset V$ and the mapping $\phi \circ f \circ \phi_V^{-1}$ is differentiable. If f is a bijection and its inverse is also differentiable, then f is called a **diffeomorphism.**

We can also define the rank $r(f)_x$ of a smooth function at x to be the rank of the above representation of f in terms of the chart. (This is independent of the choice of chart). Using this concept, we can then extend the following well-known concepts to smooth functions between manifolds. $x \in M$ is called a **regular point** of f if f has maximum rank there (i.e. if $r(f)_x = \min(\dim M, \dim N)$. In this case $y = f(x)$ is a **regular value** of f. Otherwise, x is a **singular point** of f and y is a **singular value**. (Points which are not in the range of f are also classified as regular values). An important theorem of Sard states that the set of regular values of f is dense in N.

The various versions of the inverse function theorem can then be restated without difficulty for functions between manifolds (since the claims made in these results are all local, the case of a function between differentiable manifolds can immediately be reduced to that of a function between euclidean spaces).

Proposition 6.1 *Let $f : M \in N$ be a smooth mapping between manifolds where the dimension of M is m and that of N is n. Let y be a regular value in the range of f. Then the preimage $f^{-1}(y)$ of y is an $(m-n)$-dimensional submanifold of M.*

For example, the above result implies immediately that the unit sphere is a manifold (as a submanifold of $\mathbf{R}^{n+1}$) without displaying an atlas.

Our goal in the following will be the development of the apparatus required to carry out analysis on manifolds. We will do this by beginning with local manifolds and extending to the general case with the aid of charts. We begin with the concept of the tangent space.

Definition: If U is a local manifold, then the **tangent space** to U at a point p in U is simply the product $\{p\} \times \mathbf{R}^n$. We define the **tangent bundle** over U to be the union

$$T(U) = \bigcup_{p \in U} T_p(U) = U \times \mathbf{R}^n.$$

We remark here that each **fibre** $T_p(U)$ has a natural vector space structure as a copy of $\mathbf{R}^n$.

If $f : U \to V$ is a smooth mapping between local manifolds, and if $p \in U$, then we define a mapping $(Tf)_p : T_p(U) \to T_p(V)$ by putting

$$(Tf)_p(p, x) = (f(p), (Df)_p(x)).$$

In a similar way, we define a mapping Tf from $T(U)$ into $T(V)$ by putting

$$(Tf)(p, x) = (f(p), (Df)_p(x)).$$

Roughly speaking Tf acts on the base line as f and on the fibres as the derivative of f.

In order to obtain a more intrinsic definition of the tangent bundle which can be carried over directly to the more abstract situation, we proceed as follows. We denote by $F(U)$ the set of smooth mappings from U into the real line. Then we define a **derivative** on U at the point p to be a linear mapping D from $F(U)$ into $\mathbf{R}$ so that

$$D(f \cdot g) = f(p) \cdot Dg + Df \cdot g(p)$$

for any pair of smooth functions f and g.

The typical example of a derivative is a mapping of the form $f \mapsto (Df)_p(v)$ for some $v \in V$. In fact this is the **only** example as the following result shows:

Proposition 6.2 *If D is a derivative at p and we define the vector v to be $(D\xi_1, \ldots, D\xi_n)$, then $Df = (Df)_p(v)$ for all smooth functions f. (Recall that ξ_i denotes the function which projects the vector x onto its $i-$th component).*

Proof. Without loss of generality, we assume that $p = 0$. First note that if f is a constant, then $Df = 0$. For $D(1) = D(1.1) = D(1) + D(1)$ and so $D(1) = 0$. Now we consider the formula $f(x) = f(0) + \sum \xi_i g_i$ for f near 0 (we are assuming that $p = 0$ for convenience, as we may without loss of generality). Here g_i is the smooth function

$$g_i(x) = \int_0^1 D_i f(tx) dt.$$

If we apply D to both sides, we obtain the equality

$$Df = D(f(0)) + \sum g_i(0) D\xi_i = \sum D_i f(0) v_i$$

where $v_i = D\xi_i$ which is the required result.

■

Using this fact, we can identify the tangent space $T_p(U)$ of U with the (vector space of) derivatives at p.

Note that if c is a curve in U which passes through p (say with $c(0) = p$ for the sake of simplicity), then we can regard the pair $(p, c'(0))$ as an element of the tangent space $T_p(U)$. In fact, every element (p, v) therein arises in this way as the reader can easily verify (consider the curve $c(t) = p + tv$). Hence the tangent space at p can also be regarded as the family of tangents to the curves through p (hence the name).

Note that the canonical basis for the tangent space at U can be interpreted as the set of derivatives $f \mapsto (\frac{\partial f}{\partial \xi_i})_p$ on $F(U) (i = 1, \ldots, n)$. For this reason, the basis is often written in the form $(\frac{\partial}{\partial \xi_1}, \ldots, \frac{\partial}{\partial \xi_n})$ and a typical tangent vector has the familiar representation

$$a_1 \frac{\partial}{\partial \xi_1} + \cdots + a_n \frac{\partial}{\partial \xi_n}.$$

A **vector field** on a local manifold is a smooth mapping $\tilde{\xi} : U \to T(U)$ so that $\tilde{\xi}(p) \in T_p(U)$ for each p in U. Hence the mapping has the form $p \mapsto (p, \xi(p))$ for some smooth function ξ from U into $\mathbf{R}^n$. In future, we shall not distinguish between the two mappings ξ and $\tilde{\xi}$. (We remark here that in the physics literature, the phrase "vectorüsually refers to a vector field).

A vector field $\tilde{\xi}$ induces an operator X on $F(U)$ by mapping a smooth function f into the function

$$p \mapsto (Df)_p \xi(p)$$

i.e. the value of Xf at p is the directional derivative of f along the field. This mapping is easily seen to be linear and to satisfy the property

$$X(f \cdot g) = f \cdot X(g) + g \cdot X(f).$$

In fact, every mapping X on $F(U)$ with these properties is induced in this way by a vector field. For if X is such a mapping and p is a point in U, then the function $f \mapsto (Xf)(p)$ is a derivative at p and so has the form $f \mapsto (Df)_p(\xi(p))$ for some vector $\xi(p)$ in the tangent plane $T_p(U)$. The function ξ is then a vector field which induces the operator X in the above manner.

This alternative description of vector fields is very useful and we shall shuttle back and forth between the two. The choice of symbol X or ξ will indicate which aspect we are emphasising.

Locally, a vector field has a representation

$$X = X_1 \frac{\partial}{\partial \xi_1} + \cdots + X_n \frac{\partial}{\partial \xi_n}$$

where the X_i are smooth, real-valued functions on U. This just means that

$$(X_1(p), \ldots, X_n(p))$$

is the representation of X at the point p with respect to the natural basis.

Vector fields arise as the right hand sides of first order differential equations i.e. those of the form

$$\frac{dx_i}{dt} = X_i(x) \quad (i = 1, \ldots, n).$$

A solution to such an equation is a curve c in U so that $c'(t) = X(c(t))$ for each t.

It follows from the existence theory for ordinary differential equations that for each $x_0 \in U$ and each $t_0 \in \mathbf{R}$, there is a solution c of this equation which is defined on an interval of the form $]t_0 - \eta, t_0 + \eta[$ and which is such that $c(t_0) = x_0$. Such solutions are called **integral curves** of the field.

The cotangent bundle: The cotangent bundle of a local manifold U is the union $\bigcup T_p(U)^*$ of the duals of the tangent spaces. A **one form** is a mapping ω from U into the cotangent bundle, so that $\omega(p) \in T_p(U)^*$ for each p. In other words, it assigns to each p a linear functional on the tangent space there. The standard example of a one form is the **differential** df of a smooth function f on U. This is defined by the equation $df(X_p) = X(f)_p$ i.e. at the point p, df is the linear form which associates to each tangential vector v at p the directional derivative of f in the direction v.

If we once again abuse the notation by writing ξ_i for the mapping $x \mapsto \xi_i$ so that $d\xi_i$ now denotes the derivative of this function, then $(d\xi_1, \ldots, d\xi_n)$ is the natural basis for the cotangent space (i.e. it is dual to the basis $(\frac{\partial}{\partial \xi_1}, \ldots, \frac{\partial}{\partial \xi_n})$). Hence a typical one form has the more familiar coordinate representation

$$\omega(x) = a_1(x) d\xi_1 + \cdots + a_n(x) d\xi_n.$$

More generally, we can define **k-forms** as follows: $\Lambda^k(T_p(U)$ denotes the set of alternating k-forms on $T_p(U)$. Then an **alternating k-form** or **exterior differential** on U is a mapping $\omega : U \to \bigcup_{p \in U} \Lambda^k(T_p(U)$ so that $\omega(p) \in \Lambda^k(T_p(U))$ for each p. Since $(d\xi_1, \ldots, d\xi_n)$ is a basis for the cotangent space, we have the basis

$$\{d\xi_{i_1} \wedge \cdots \wedge d\xi_{i_k} : i_1 < i_2 < \cdots < i_k\}$$

for the space of k-forms on the tangent space. Thus each k-form has a representation

$$\omega = \sum_{1 \leq i_1 < \cdots < i_k \leq n} a_{i_1 \ldots i_k} d\xi_{i_1} \wedge \cdots \wedge d\xi_{i_k}.$$

If ω is such a form, we define its **differential** as follows:

$$d\omega = \sum_{1\le i_1<\cdots<i_k\le n} da_{i_1\ldots i_k}\wedge d\xi_{i_1}\wedge\cdots\wedge d\xi_{i_k} \tag{44}$$

$$= \sum_{1\le i_1<\cdots<i_k\le n}\sum_{i=1}^{n}\frac{\partial}{\partial\xi_i}a_{i_1\ldots i_k}d\xi_i\wedge d\xi_{i_1}\wedge\cdots\wedge d\xi_{i_k}. \tag{45}$$

We note some simple properties of this operator. The proofs are routine calculations.
a) $d(\omega\wedge\eta) = d\omega\wedge\eta + (-1)^{kl}(\omega\wedge d\eta)$ where ω is a k-form, η an l-form;
b) $dd\omega = 0$.

Examples:
$n = 2, k = 1$. A typical one-form on the plane has the form

$$\omega = a_1(x,y)dx + a_2(x,y)dy.$$

Then

$$d\omega = (\frac{\partial a_2}{\partial x} - \frac{\partial a_1}{\partial y})dx\wedge dy.$$

$n = 3, k = 1$. A typical one-form in space has the form

$$\omega = a_1dx + a_2dy + a_3dz$$

where the a's are smooth functions. Then

$$d\omega = \Big(\frac{\partial a_3}{\partial y} - \frac{\partial a_2}{\partial z}\Big)dy\wedge dz + \Big(\frac{\partial a_1}{\partial z} - \frac{\partial a_3}{\partial x}\Big)dz\wedge dx + \Big(\frac{\partial a_2}{\partial x} - \frac{\partial a_1}{\partial y}\Big)dx\wedge dy$$

(cf. the curl of the vector function (a_1, a_2, a_3)).
$n = 3, k = 2$. If

$$\omega = a_1dy\wedge dz + a_2dz\wedge dx + a_3dx\wedge dy$$

then

$$d\omega = \Big(\frac{\partial a_1}{\partial x} + \frac{\partial a_2}{\partial y} + \frac{\partial a_3}{\partial z}\Big)dx\wedge dy\wedge dz.$$

(c.f. the divergence of the vector function (a_1, a_2, a_3)).

In order to formulate the above concepts within the framework of general manifolds, we must work out the consequences of coordinate changes on the representation of forms. The appropriate notion is that of composition of forms with functions.

If $F : U \to V$ is a smooth mapping with components $(f_1, \ldots, f_m)$ where U is open in $\mathbf{R}^n$ and V is open in $\mathbf{R}^m$, and

$$\omega = \sum a_{i_1\ldots i_k}d\eta_{i_1}\wedge\cdots\wedge d\eta_{i_k},$$

then $\omega \circ F$ is the form:

$$\sum (a_{i_1 \ldots i_k}) \circ F df_{i_1} \wedge \cdots \wedge df_{i_k} = \sum (a_{i_1 \ldots i_k} \circ F) \left(\sum_{j_1=1}^{n} \partial f_{i_1} \partial \xi_{j_1} d\xi_{j_1} \right) \wedge \cdots \wedge \left(\sum_{j_k=1}^{n} \partial f_{i_k} \partial \xi_{j_k} d\xi_{j_k} \right)$$

Example: If ω is the form $d\eta_i$, then

$$(d\eta_i) \circ F = \sum_{j=1}^{n} \frac{\partial f_i}{\partial \xi_j} d\xi_j = df_i.$$

The following simple properties can be verified by means of routine calculations:
a) $(\omega_1 + \omega_2) \circ F = \omega_1 \circ F + \omega_2 \circ F$;
b) if g is a smooth function, then $(g\omega) \circ F = (g \circ F)(\omega \circ F)$;
c) $(\omega \wedge \eta) \circ F = (\omega \circ F) \wedge (\eta \circ F)$;
d) $(d\omega) \circ F = d(\omega \circ F)$;
e) $(\omega \circ F) \circ G = \omega \circ (F \circ G)$.

Suppose that F is a smooth function on $\mathbf{R}^n$ and that ω is an n-form on the same space, say $\omega = a d\xi_1 \wedge \cdots \wedge d\xi_n$. Then

$$\omega \circ F = a \circ F df_1 \wedge \cdots \wedge df_n = a \circ F \det(DF) d\xi_1 \wedge \cdots \wedge d\xi_n.$$

Integration of k-forms: If ω is a k-form on an open subset U of $\mathbf{R}^k$, then it has a representation $a d\xi_1 \wedge \cdots \wedge d\xi_k$ for some smooth function a on U. We say that ω is integrable if a is and define

$$\int_U \omega = \int_U a(\xi_1, \ldots, \xi_k) d\xi_1 \ldots d\xi_k.$$

It follows immediately from the definition and the transformation law for integrals, that if $F : U \to V$ is a diffeomorphism so that the determinant of its Jacobi matrix is always positive, then $\int_U \omega = \int_V \omega \circ F$.

If we combine the definitions of compositions of forms with functions and the above integral, we obtain a definition of integration of k-forms on $\mathbf{R}^n$ along k-dimensional submanifolds which simultaneously generalises line integrals and surface integrals etc.

Definition: A (singular) **parametrised k-cube** in $U \subset \mathbf{R}^n$ is a smooth mapping
$c : I^k \to U$. As in the case of a curve, we shall identify two such cubes c and c' if $c = c' \circ F$ where $F : I^k \to I^k$ is a diffeomorphism whose Jacobi matrix has positive determinant everywhere. Then if ω is a k-form on U, we define $\int_c \omega = \int_{I^k} \omega \circ c$. This is independent of the parametrisation of the singular cube.

More generally, we define a **k-chain** to be an expression of the form $c = n_1 c_1 + \cdots + n_r c_r$ where each n_i is an integer and each c_i is a k-cube. Then we define

$$\int_c \omega = n_1 \int_{c_1} \omega + \cdots + n_r \int_{c_r} \omega.$$

We now define the **boundary** ∂c of a k-cube to be the $(k-1)$-chain constructed as follows: for each $i = 1, \ldots, k$ consider the mappings

$$I_i^b : (\xi_1, \ldots, \xi_{k-1}) \mapsto (\xi_1, \ldots, \xi_{i-1}, 0, \xi_i, \ldots, \xi_{k-1})$$

and

$$I_i^t = (\xi_1, \ldots, \xi_{k-1}) \mapsto (\xi_1, \ldots, \xi_{i-1}, 1, \xi_1, \ldots, \xi_{k-1}).$$

Then ∂c is the chain

$$\sum (-1)^{i-1}(c \circ I_i^t - c \circ I_i^b).$$

We are now in a position to state a very general form of Stokes' theorem.

Theorem 6.3 *If ω is a $(k-1)$-form on a k-cube c, then $\int_{\partial c} \omega = \int_c d\omega$.*

Proof. We prove the result initially for the case where c *is* the unit cube i.e. the mapping c is the identity. ω has the form

$$a_1(x) d\xi_2 \wedge \cdots \wedge d\xi_k + \cdots + a_k(x) d\xi_1 \wedge \cdots \wedge d\xi_{k-1}$$

and it suffices to prove the result for each term, say the first one. Then

$$d\omega = \frac{\partial a_1}{\partial \xi_1} d\xi_1 \wedge \cdots \wedge d\xi_k$$

and

$$\begin{aligned} \int_{I^k} d\omega &= \int_{I^k} \frac{\partial a_1}{\partial \xi_1} d\xi_1 d\xi_2 \ldots d\xi_k \\ &= \int_{I^{k-1}} (\int_0^1 \frac{\partial a_1}{\partial \xi_1} d\xi_1) d\xi_2 \ldots d\xi_k \\ = \int_{I^{k-1}} a_1(1, \xi_2, \ldots, \xi_k) d\xi_2 \ldots d\xi_k &- \int_{I^{k-1}} a_1(0, \xi_2, \ldots, \xi_k) d\xi_2 \ldots d\xi_k. \end{aligned}$$

But

$$\omega \circ I_1^t = a_1(1, \xi_1, \ldots, \xi_{k-1}) d\xi_1 \wedge \cdots \wedge d\xi_{k-1}$$

and

$$\omega \circ I_1^b = a_1(0, \xi_1, \ldots, \xi_{k-1}) d\xi_1 \wedge \cdots \wedge d\xi_{k-1}$$

and for each other i, $\omega \circ I_i^t = \omega \circ I_i^b = 0$. Hence

$$\int_{I^k} d\omega = \int_{I^k} (\omega \circ I_1^t - \omega \circ I_1^b) + 0 = \int_{\partial I^k} \omega.$$

In the case of a general cube, we have

$$\int_c d\omega = \int_{I^k} (d\omega) \circ c = \int_{I^k} d(\omega \circ c) = \int_{\partial I^k} \omega \circ c = \int_{\partial c} \omega.$$

.

■

We now discuss these concepts within the framework of an abstract manifold. We begin with the tangent space $T_p(M)$. This can be defined in three ways:
1. as the set of all derivatives at p of $F(M)$;
2. as the set of equivalence classes of triples (U, ϕ, x) where U is a chart that contains p and is such that $\phi(p) = 0$ and x is a vector in $\mathbf{R}^n$. Two such triples (U, ϕ, x) and (U_1, ϕ_1, x_1) are **equivalent** if $x_1 = D(\phi_1 \circ \phi^{-1})(x)$;
3. as the set of equivalence classes of curves $c : I \to M$ where I is an interval about 0 and $c(0) = p$. The equivalence relation is that one which identifies c and c_1 whenever $(\phi \circ c_1)'(0) = (\phi \circ c_2)'(0)$ for any chart.

If (U, ϕ, t) and (V, ψ, t') are two representants of a tangent vector at $p \in M$, and if we write ϕ and ψ in coordinate form $\phi = (x_1, \ldots, x_n)$, $\psi = (x'_1, \ldots, x'_n)$, then the relation $t' = D(\psi \circ \phi^{-1})_{\phi(p)}(t)$ becomes

$$t'_i = \sum_j t_j \frac{\partial x'_i}{\partial x_j}$$

where $t' = (t'_1, \ldots, t'_n)$, $t = (t_1, \ldots, t_n)$ and we have written, with an abuse of notation, $\frac{\partial x'_i}{\partial x_j}$ for the $(i, j)-$th element of the Jacobi matrix of $\psi \circ \phi^{-1}$ i.e. $D_j(x_i \circ \phi^{-1})$.

The **tangent bundle** $T(M)$ is the union $\bigcup_p T_p(M)$ of the tangent spaces at the points of M.

If $\phi : M \to N$ is a smooth mapping between manifolds, then ϕ induces a mapping $f \mapsto f \circ \phi$ from $F(N)$ into $F(M)$ and so, by duality, a mapping, which we shall denote by $T(\phi)$ from $T_p(N)$ into $T_{\phi(p)}(N)$. It can also be constructed as follows: if v is a vector in $T_p(M)$, then it corresponds to the tangent of a curve c on M. $T_\phi(v)$ is defined to be the tangent of the image curve $\phi \circ c$.

The **cotangent bundle** of a manifold M is the set $T^*(M) = \bigcup_{p \in M} T_p(M)^*$. A **one-form** on M is then a smooth mapping

$$\omega : M \to T^*(M)$$

such that $\omega(p) \in T_p(M)^*$ for each p. Similarly, a **k-form** is defined to be a mapping ω from M into the bundle $\Lambda^k(T(M))$ so that $\omega(p) \in \Lambda^k(T_p(M))$ for each p.

A **vector field** on a manifold can be regarded as a mapping ξ from M into the tangent bundle such that $\xi(p) \in T_p(M)$ for each p. Equivalently, it is a linear mapping X from $F(M)$ into $F(M)$ so that $X(fg) = fX(g) + gX(f)$ for $f, g \in F(M)$. We write $X(M)$ for the set of all vector fields on M.

A vector field induces a local flow on a manifold i.e. for each $t_0 \in \mathbf{R}$, there is a positive ϵ so that for each $p \in M$ there is a smooth curve c defined on the interval $]t_0 - \epsilon, t_0 + \epsilon[$ on the manifold with $c(t_0) = p$ and, for each t in the interval of definition, $c'(t) = \xi_{c(t)}$. This is proved by using a coordinate chart to translate this into a differential equation in n-dimensional euclidean space whose solution,

which is guaranteed by standard results on such equations, is then transferred back to the manifold, again using the chart.

We would now like to be able to integrate k-forms on k-dimensional manifolds. In order to do this we require the following concept: an **oriented manifold** is a manifold M with an atlas $\{(U_\alpha, \phi_\alpha)\}$ so that for each pair α and β, the transformation function $\phi_\beta \circ \phi_\alpha^{-1}$ is an orientation preserving mapping on n-space (i.e. the determinant of its Jacobi matrix is always positive).

Now consider an n-form ω on an oriented manifold M. We wish to define its integral $\int_M \omega$. In order to do this we suppose initially that ω is supported by a chart (U, ϕ) (i.e. that $\omega = 0$ on $M \backslash U$). We then define $\int_M \omega$ to be $\int_{\phi(U)} \omega \circ \phi^{-1}$. This is well-defined (i.e. independent of the choice of chart), a fact which follows from the formula for the change of variable in integrals (note that at this point it is important that the manifold be oriented).

In order to dispense with the condition that ω be supported by a chart, we use a standard technique for passing from local concepts defined via charts to global ones. This is the use of so-called **partitions of unity**. Recall that if U is an open cover of a topological space, then a **locally finite refinement** of U is a refinement V with the property that each point x has a neighbourhood V which meets only finite many $U \in V$. We shall require the fact that metric spaces have the property that each open covering has a locally finite refinement. In particular, manifolds have this property. If we start with a covering of a manifold by charts, then the open sets of such a refinement are also charts. It is then easy to see that one can find a **smooth partition of unity subordinate to** U i.e. a family $\{\phi_U\}$ of smooth functions from M into $[0, 1]$ indexed by the original atlas, such that

a. ϕ_U has its support in U;

b. $\{\phi_U\}$ is locally finite;

c. the sum of the ϕ_U is 1.

We can now extend the definition of the integral of a form to the general situation as follows: if (ϕ_α) is a partition of unity subordinate to a chart of M, then $\int_M \phi_\alpha \omega$ is defined for each α. We can then define $\int_M \omega$ simply to be $\sum \int_{U_\alpha} \phi_\alpha \omega$ provided that this sum converges (in which case the form is said to be **integrable**). This is certainly the case if the form has compact support.

In order to give a general form of Stokes' theorem, we require the concept of a **manifold with boundary**. This is defined to be a topological manifold with a chart $\{(U_\alpha, \phi_\alpha)\}$ where now the range of each ϕ_α is either $\mathbf{R}^n$ or

$$H_+^n = \{x \in \mathbf{R}^n : \xi_1 \geq 0\}.$$

Once again, we demand that the transformation functions between two charts be smooth. The **boundary** ∂M of M is then defined to be the set of all those points which are mapped into the boundary of H_+^n by one (and hence by any) chart. This is a manifold with dimension one less than that of M. We can define the

concept of smooth real-valued function on a manifold with boundary, respectively of a smooth function between manifolds with boundary in the natural way. Also if M is an oriented manifold, there is a natural way to orient the boundary. We can now state without proof a version of Stokes' theorem in this context.

Theorem 6.4 *Suppose that ω is an n-form on an n-dimensional oriented manifold M with boundary. Then $\int_M d\omega = \int_{\partial M} \omega$.*

6.1 Riemann manifolds

These bear the same relationship to differentiable manifolds as euclidean spaces do to affine spaces. The typical examples are **hypersurfaces** in $\mathbf{R}^n$. These can be defined locally either

as subsets of the form $M = \{x \in V : f(x) = 0\}$ where V is open in $\mathbf{R}^n$ and f is a smooth real-valued function on V whose gradient $(Df)_x$ never vanishes on M;

or

as the image of a smooth mapping ϕ from an open subset U of $\mathbf{R}^{n-1}$ into $\mathbf{R}^n$ whose Jacobi matrix has maximum rank (i.e. $n-1$) at each point of U. In the latter case, the vectors $(D_1\phi, \ldots, D_{n-1}\phi)_{(x)}$ span an $(n-1)$-dimensional subspace of $\mathbf{R}^n$ - the **tangent space** $T_p(M)$ of M at $p = \phi(x)$.

The standard inner product on $\mathbf{R}^n$ induces one on the tangent space. This has coefficients $g_{ij} = (\phi_i|\phi_j)$ with respect to above basis. The form with coefficients (g_{ij}) is called the **first fundamental form** or **Riemannian metric** of the surface.

A Gauß mapping is then a mapping $x \mapsto \mathbf{N}(x)$ on U with the property that $|\mathbf{N}(x)| = 1$ and $\mathbf{N}(x) \perp T_p(M)$ for each x (whereby $p = \phi(x)$). If M has the local description

$$\{p : F(p) = 0\}$$

, then

$$\mathbf{N}(x) = \frac{\operatorname{grad} F(p)}{|\operatorname{grad} F(p)|}$$

is a suitable Gauß mapping.

Examples of hypersurfaces:
I. Hyperplanes i.e. sets of the form $\{x : f(x) = \alpha\}$ where f is a non-zero linear form on $\mathbf{R}^n$.
II. Landscapes i.e. surfaces of the form

$$\xi_n = f(\xi_1, \ldots, \xi_{n-1})$$

where f is a smooth mapping.
III. The unit sphere $S^{n-1} = \{x \in \mathbf{R}^n : |x| = 1\}$.

IV. Cylinders i.e. surfaces of the form

$$\{x : f(\xi_1, \ldots, \xi_{n-1}) = 0\}$$

for suitable smooth functions f on $\mathbf{R}^{n-1}$.

If M is a parametrised surface, then the Gauß mapping is the function

$$\mathbf{N} : u \mapsto \frac{\phi_1(u) \times \cdots \times \phi_{n-1}(u)}{|\phi_1(u) \times \cdots \times \phi_{n-1}(u)|}.$$

(We are using the vector product in n-dimensional space which is a function of $(n-1)$ arguments - see Lineare Algebra).

The **second fundamental form** II_u is the bilinear form on the tangential space which has matrix $[l_{ij}]$ with respect to the natural basis for $T_p(M)$. Here the entries of the matrix are defined by the equations

$$l_{ij} = (\phi_i | \mathbf{N}_j) = -(\phi_{ij} | \mathbf{N}).$$

II_u is determined by a self-adjoint mapping L_p on the tangent space. This means that

$$II_u(x, y) = (L_p x | y).$$

L_p has the matrix $[g_{ij}]^{-1}[l_{ij}]$ with respect to the basis $\phi_1, \ldots, \phi_{n-1}$. In fact, L_p is the mapping $(-D\mathbf{N}) \circ (D\phi)^{-1}$ from $T_p(M)$ into $T_p(M)$.

Using L_p, we define the k-th fundamental form on the tangent space by means of the formula

$$(x, y) \mapsto (L_p^{k-1} x | y).$$

Note that in three dimensions we have the relationship

$$L_p^2 - H(p) L_p + \kappa(p) Id = 0$$

(Cayley-Hamilton theorem) which implies the relationship

$$III_p - H(p) II_p + \kappa(p) I_p = 0$$

between the first three fundamental forms.

Using the fundamental forms, we can define the following concepts as in the case of two-surfaces:

the Gaussian curvature $\kappa(p) = \det L_p$;

the mean curvature $H(p) = \operatorname{Tr} L_p$;

the principal curvatures i.e. the eigenvalues of L_p;

the principal directions i.e. the eigenvectors of L_p;

lines of curvature i.e. curves c on the surface so that $c'(t)$ is an eigenvector of $L_{c(t)}$ for each t;

umbilical points i.e. points where L_p is a multiple of the identity;

conjugate tangent vectors i.e. vectors x and y in $T_p(M)$ with $(L_p x|y) = 0$;

asymptotic directions i.e. tangent vectors $x \in T_p(M)$ with $(L_p x|x) = 0$;

the Dupin indicatrix i.e. the subset $\{x : (L_p x|x) = \pm 1\}$ of the tangent space. This is a conic section whose form describes the nature of the surface near p. We now turn to the topic of **covariant differentiation**. We begin with the local situation. Suppose that X and Y are vector fields on $U \subset \mathbf{R}^n$. Then we define the field $\nabla_X Y$ by the formula

$$\nabla_X(Y)(p) = (dY(p))X(p).$$

(i.e. we calculate the derivative of Y in the direction of X). In coordinates, this has the form

$$\nabla_X Y = \sum_j (\sum_i X_i \frac{\partial Y_j}{\partial \xi_i}) \frac{\partial}{\partial \xi_j}.$$

The following simple properties of this differentiation can be verified directly:

$\nabla_X Y$ is linear in X and Y;

$\nabla_X(fY) = (Xf)Y + f\nabla_X Y$ $(f \in F(U))$;

$\nabla_X Y - \nabla_Y X = [X, Y]$;

$\nabla_X((Y|Z)) = (\nabla_X Y|Z) + (Y|\nabla_X Z)$.

For a general manifold, there is no analogue of the notion of a covariant derivative. Hence we introduce the following concept axiomatically. Later we shall see that the presence of a Riemannian structure, for example, ensures the existence of such a differentiation.

Definition Let M be a manifold. A **connection** on M is a mapping which assigns to each vector field X on M a linear operator ∇_X on $X(M)$ so that
1) $\nabla_X(fY) = X(f)Y + f\nabla_X(Y)$;
2) $\nabla_{fX+gX_1} = f\nabla_X + g\nabla_{X_1}$.

Example There is a natural connection on a hypersurface $M \subset \mathbf{R}^{n+1}$ which is defined as follows: let $\bar{\nabla}_X(Y)$ be the usual connection on $\mathbf{R}^{n+1}$. Then if X and Y are two vector fields on M, we define $(\nabla_X(Y))_p$ to be the orthogonal projection of $\bar{\nabla}_X(Y)$ onto the tangent space $T_p(M)$ i.e.

$$(\nabla_X(Y))_p = \bar{\nabla}_X(Y) - (\bar{\nabla}_X(Y)|\mathbf{N}(p))\mathbf{N}(p).$$

In order to describe a connection on a manifold in coordinates, we consider two vector fields X and Y. Suppose that (U, ϕ) is a chart and that X_U and Y_U are the representations of X and Y in $\phi(U)$. The connection $\nabla_X(Y)$ of X and Y corresponds to a connection $\tilde{\nabla}_{X_U}(Y_U)$ in $\mathbf{R}^n$. Of course, this is not, in general,

the natural one defined above. It is described by the Christoffel symbols of the connection which are defined as follows:

Definition: Let $\nabla_X Y$ be a connection in n-space. Then the Christoffel symbols are the functions Γ_{ij}^k which are defined by the equation:

$$\nabla_{\frac{\partial}{\partial \xi_i}}(\frac{\partial}{\partial \xi_j}) = \sum_k \Gamma_{ij}^k \frac{\partial}{\partial \xi_k}$$

Thus the connection can be expressed locally in the form

$$\nabla_X(Y) = \tilde{\nabla}_X Y + \sum_k (\sum_{i,j} \Gamma_{ij}^k X_i Y_j) \frac{\partial}{\partial \xi_k}$$

where

$$X = \sum_{i=1}^n X_i \frac{\partial}{\partial \xi_i}$$

and $\tilde{\nabla}_X Y$ is the standard connection on $\mathbf{R}^n$. (Warning: the Christoffel symbols are not 3-tensors. Their behaviour under coordinate changes is more complicated).

Example: If M is a hypersurface in $\mathbf{R}^n$ with parametrisation ϕ so that the first fundamental form has matrix $G = [g_{ij}] = [(\phi_i|\phi_j)]$, then the Christoffel symbols of the connection described above are given by the formulae:

$$\Gamma_{ijk} = \frac{1}{2}(\frac{\partial g_{jk}}{\partial \xi_i} + \frac{\partial g_{ki}}{\partial \xi_j} - \frac{\partial g_{ij}}{\partial \xi_k})$$

resp.

$$\Gamma_{ij}^k = \sum_l \Gamma_{ijl} g^{lk}$$

where $[g^{lk}]$ is the inverse of G.

We now return to the general setting of a manifold M with connection. Then we define the **torsion field** K by putting

$$K(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]$$

and the **Riemann curvature field** R by

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_X \nabla_Y Z - \nabla_{[X,Y]} Z.$$

K and R are tensors which satisfy the following identities:

$$K(X,Y) = -K(Y,X)$$

i.e. K is anti-symmetric.

$$R(X,Y)Z = -R(X,Y)Z$$

i.e. R is anti-symmetric in X and Y.

Further if K vanishes identically (which is the case for a Riemannian manifold as we shall see below), then we have the identity

$$R(X,Y)Z + R(Z,X)Y + R(Y,Z)X = 0.$$

In local coordinates, K is the 3-tensor with coefficients $(\Gamma^k_{ij} - \Gamma^k_{ji})$ (so that the condition $K = 0$ means that the symbol Γ^k_{ij} is symmetric in i and j). R is the tensor with coordinates

$$R^k_{lij} = \Gamma^k_{is} - \sum_s \Gamma^k_{js}\Gamma^s_{il} + \frac{\partial \Gamma^k_{jl}}{\partial \xi_i} - \frac{\partial \Gamma^k_{il}}{\partial \xi_j}.$$

Definition: A **local Riemannian manifold** is a local manifold i.e. an open subset U of an $\mathbf{R}^n$, together with a **metric tensor** i.e. n^2 smooth mappings g_{ij} from U into $\mathbf{R}$ so that for each $u \in U$, the matrix $[g_{ij}(u)]$ is positive definite. There is a corresponding global definition i.e. a **Riemann manifold** is a differentiable manifold so that each chart is assigned a metric tensor in a compatible way. Since for our examples the concept of a local manifold suffices, we shall not go into the formal definition. In any case, even for manifolds which are not local actual calculations take place via charts i.e. within the context of a local manifold.

A more intrinsic definition of ∇ can be given as follows: we define $\nabla_X Y$ for two vector fields X, Y by specifying that the following equation holds for each vector field Z:

$$2g(\nabla_X Y, Z) = X(g(Y,Z)) + Y(g(Z,X)) - Z(g(X,Y)) +$$
$$g(Z,[X,Y]) + g(Y,[Z,X]) - g(X,[Y,Z]).$$

Using this connection, the torsion field and Riemann tensor resp. the notion of a stationary vector field, parallel transform and geodetics are defined for a Riemann manifold.

In applications, one of the main purposes of a connection is to connect two tangent planes to a manifold at distinct points by means of a suitable curve. This is done as follows. Suppose that $\tilde{c}$ is a smooth function on M with $\tilde{c}(a) = p$ and $\tilde{c}(b) = q$. A **vector field along** $\tilde{c}$ is a smooth function $X : [a,b] \to T(M)$ so that for each $t \in [a,b]$, $X(t) \in T_{\tilde{c}(t)}(M)$.

A typical example of such a field is the tangent field of $\tilde{c}$ i.e. the case where $X(t) = \dot{\tilde{c}}(t)$.

Such a vector field is said to be **stationary** along $\tilde{c}$ if its derivative with respect to the tangent field is zero i.e. if $\nabla_{\dot{\tilde{c}}(t)}(X(t)) = 0$ for each t. In coordinates, this corresponds to the differential equation

$$\frac{dX_k}{dt} + \sum_{i,j} \Gamma^k_{ij}\dot{c}_i(t)X_j(t) = 0$$

for $k = 1, \ldots, n$. This is a linear system of first order equations and it follows from the theory of such equations that for each vector x in the tangent space at p, there is a unique vector $y = P_{\tilde{c}}(x)$ in $T_q(M)$ so that there is a stationary vector field X along $\tilde{c}$ with $X(a) = x$ and $X(b) = y$. The mapping $x \mapsto P_{\tilde{c}}(x)$, which is an isomorphism from $T_p(M)$ onto $T_q(M)$, is called **parallel translation** along $\tilde{c}$.

We remark that in the case of a Riemann manifold, parallel translation is an isometry between the corresponding tangent spaces.

In the case of a Riemann manifold, we can define the **length** of a curve c as follows:

$$L(c) = \int_a^b \sqrt{g_{c(t)}(\dot{c}(t), \dot{c}(t))} dt.$$

The **energy** of c is

$$E(c) = \frac{1}{2} \int_a^b g_{c(t)}(\dot{c}(t), \dot{c}(t)) dt.$$

We have the following simple relation between these quantities:

$$L(c)^2 \leq 2E(c)|I|$$

(where $|I|$ is the length of the domain I of definition of c). We have equality in the above inequality if and only if the speed $|\dot{c}(t)|$ of the curve is constant. The geodetics of a surface can be characterised by suitable extremal properties of these functionals.

If M is a Riemann manifold and p, q are points of M, we define

$$\tilde{d}(p, q) = \inf\{L(c) : c(0) = p, c(1) = q\}.$$

Note that if g is a Riemann metric on an open subset of $\mathbf{R}^n$ and K is compact in U, then there are constants M and m so that

$$md(p, q) \leq \tilde{d}(p, q) \leq Md(p, q)$$

for $p, q \in K$. It follows from this that the above metric on a Riemann manifold induces the original topology.

We now state without proof a characterisation of manifolds which have the property that the geodetic equations have global solutions (i.e. solutions which are defined on the whole of $\mathbf{R}$).

The Hopf-Rinov theorem: For a Riemann manifold, the following are equivalent:
a) M is complete (i.e. geodesics are infinitely extendable):
b) bounded, closed subsets of M are compact:
c) M is complete under the metric $\tilde{d}$;
d) there is a point on the manifold from which all geodesics are extendable.
If any of these conditions are satisfied and the manifold is connected, then any two points on the manifold can be joined by a curve of shortest length.

In particular, each of these conditions is satisfied if the manifold is compact.