I. Classical Differential Geometry of Curves

This is a first course on the differential geometry of curves and surfaces. It begins with topics mentioned briefly in ordinary and multivariable calculus courses, and two major goals are to formulate the mathematical concept(s) of curvature for a surface and to interpret curvature for several basic examples of surfaces that arise in multivariable calculus.

Basic references for the course

An obvious starting point is to give the official text for the course:

M. Lipschutz, *Schaum's Outlines – Differential Geometry*, Schaum's/McGraw-Hill, 1969, ISBN 0–07–037985–8.

This is actually a review book on differential geometry, but it contains a great deal of information on the classical approach, brief outlines of the underlying theory, and many worked out examples.

These notes are intended to expand upon the content of the text and, to some extent, reflect the content of the lectures. The following items are similar in spirit to the course:

C. Baer, *Elementary Differential Geometry*, Cambridge Univ. Press, New York, 2010, ISBN 978–0–521–89671–9.

T. Shifrin, *Differential Geometry: A First Course on Curves and Surfaces*, freely available online: `http://www.math.uga.edu/≃shifrin/ShifrinDiffGeo.pdf`

P. A. Blaga, *Lectures on the Differential Geometry of Curves and Surfaces*, Napoca Press, Cluj-Napoca, Romania, 2005, ISBN 9736568962.

R. S. Millman and G. D. Parker, *Elements of Differential Geometry*, Prentice-Hall, Englewood Cliffs, NJ, 1977, ISBN 0-13-264243-7.

At various points we shall also refer to the following alternate sources, which are texts at slightly higher levels:

J. Oprea, *Differential Geometry and Its Applications* (Second Ed.), Mathematical Association of America, Washington, DC, 2006, ISBN 978–0–88385–748–9.

A. Pressley, *Elementary Differential Geometry*, Springer-Verlag, New York NY, 2000, ISBN 978–1852331528.

M. P. do Carmo, *Differential Geometry of Curves and Surfaces*, Prentice-Hall, Saddle River NJ, 1976, ISBN 0–132–12589–7.

J. A. Thorpe, *Elementary Topics in Differential Geometry*, Springer-Verlag, New York, 1979, ISBN 0–387–90357–7.

J. J. Stoker, *Differential Geometry*, Wiley, New York, 1949, ISBN 0–471–50403–3.

N. J. Hicks, Notes on differential geometry (Van Nostrand Mathematical Studies No. 3). *D. Van Nostrand, New York*, 1965.
(Online: `http://www.wisdom.weizmann.ac.il/∼yakov/scanlib/hicks.pdf`)

W. Kühnel, Differential Geometry: Curves – Surfaces – Manifolds (Student Mathematical Library, Vol. 16, Second Edition, transl. by B. Hunt). *American Mathematical Society, Providence, RI*, 2006. **ISBN-10:** 0-8218-3988-8.

B. O'Neill, *Elementary Differential Geometry.* (Revised Second Edition), Elsevier/Academic Press, San Diego CA, 2006, ISBN 0–12–088735–5.

Mathematical prerequisites

At many points we assume material covered in previous mathematics courses, so we shall include a few words on such background material. This course explicitly assumes prior experience with the elements of linear algebra (including matrices, dot products and determinants), the portions of multivariable calculus involving partial differentiation, and some familiarity with the a few basic ideas from set theory such as unions and intersections. At a few points in later units we shall also assume some familiarity with multiple integration. but we shall not be using results like Green's Theorem, Stokes' Theorem or the Divergence Theorem. For the sake of completeness, files describing the background material (with references to standard texts that have been used in the Department's courses) are included in the course directory and can be found in the files called `background∗.pdf`, where $n = 1, 2$ or 3.

How the prerequisites relate to this course

The name "differential geometry" suggests a subject which uses ideas from calculus to obtain geometrical information about curves and surfaces; since vector algebra plays a crucial role in modern work on geometry, the subject also makes extensive use of material from linear algebra. At many points it will be necessary to work with topics from the prerequisites in a more sophisticated manner, and it is also necessary to be more careful in our logic to make sure that our formulas and conclusions are accurate. Also, at numerous steps it might be necessary to go back and review things from earlier courses, and in some cases it will be important to understand things in more depth than one needs to get through ordinary calculus, multivariable calculus or matrix algebra. Frequently one of the benefits of a mathematics course is that it sharpens one's understanding and mastery of earlier material, and differential geometry certainly provides many opportunities of this sort.

The origins of differential geometry

The paragraph below gives a very brief summary of the developments which led to the emergence of differential geometry as a subject in its own right by the beginning of the 19^{th} century. Further information may be found in any of several books on the history of mathematics.

Straight lines and circles have been central objects in geometry ever since its beginnings. During the 5^{th} century B.C.E., Greek geometers began to study more general curves, most notably the ellipse, hyperbola and parabola but also other examples (for example, the Quadratrix of Hippias, which allows one to solve classical Greek construction problems that cannot be answered by means of straightedge and compass, and the Spiral of Archimedes, which is given in polar coordinates by the equation $r = \theta$). In the following centuries Greek mathematicians discovered a large number of other curves and investigated the properties of such curves in considerable detail for a variety of reasons. By the end of the Middle Ages in the 15$^{\text{th}}$ century, scientists and mathematicians had discovered further examples of curves that arise in various natural contexts, and still further examples and results were discovered during the 16$^{\text{th}}$ century. Problems involving curves played an important role in the development of analytic geometry and calculus during the 17$^{\text{th}}$ and 18$^{\text{th}}$ centuries, and these subjects in turn yielded powerful new techniques for analyzing curves and analyzing their properties. In particular, these advances created a unified framework for understanding the work of the Greek geometers and a setting for studying new classes of curves and problems beyond the reach of classical Greek geometry. Interactions with physics played a major role in the mathematical study of curves beginning in the 15$^{\text{th}}$ century, largely because curves provided a means for analyzing the motion of physical objects. By the beginning of the 19$^{\text{th}}$ century, the differential geometry of curves and surfaces had begun to emerge as a subject in its own right.

This unit describes the classical nineteenth century theory of curves in the plane and 3-dimensional space. Subsequent developments have led to more abstract and broadly based formulations of both subjects. Treatments along these lines appear in most of the books listed above. Both the classical and the more modern approaches have advantages. The classical approach usually provides the fastest way of getting to the basic properties of curves and surfaces in differential geometry and working with fundamental classes of examples, while the various modern approaches generally yield more conceptual insight into the nature of these properties.

References for examples

Here are some web links to sites with pictures and written discussions of many curves that mathematicians have studied during the past 2500 years, including the examples mentioned above:

`http://www-gap.dcs.st-and.ac.uk/∼history/Curves/Curves.html`

`http://www.xahlee.org/SpecialPlaneCurves_dir/specialPlaneCurves.html`

`http://facstaff.bloomu.edu/skokoska/curves.pdf`

Clickable links to these sites — and others mentioned in these notes — are in the course directory file `dg2012-links.pdf`.

REFERENCES FOR RESULTS ON CURVES FROM CLASSICAL GREEK GEOMETRY. A survey of curves in classical Greek geometry is beyond the scope of these notes, but here are references for Archimedes' paper on the spiral named after him and a description of the work of Apollonius of Perga (*c.* 262–*c.* 190 B.C.E.) on conic sections in (relatively) modern language.

> Archimedes of Syracuse (*author*) and T. L. Heath (*translator*), *The Works of Archimedes* (Reprinted from the 1912 Edition), Dover, New York, NY, 2002, ISBN 0–486–42084–1. (The paper *On spirals* appears on pages 151–188).

> H. G. Zeuthen, *Die Lehre von den Kegelschnitten im Altertum* (The study of the conic sections in antiquity; translation from Danish into German by R. von Fischer-Benzon), A. F. Höst & Son, Copenhagen, DK, 1886. — See the file `zeuthen.pdf` in the course directory for an online copy.

Here are links to more modern and less formal historical discussions of classical work on curves and surfaces.

http://www.ms.uky.edu/∼carl/ma330/hippias/hippias2.html

http://en.wikipedia.org/wiki/On_Spirals

http://math.ucr.edu/∼res/math153/history04X.pdf

http://math.ucr.edu/∼res/math153/history04b.pdf

Finally, here are a few more online references, some of which are cited at various points in these notes:

http://people.math.gatech.edu/∼ghomi/LectureNotes/index.html

http://en.wikipedia.org/wiki/Differential_geometry_of_surfaces

http://www.seas.upenn.edu/∼cis70005/cis700sl6pdf.pdf

http://www.math.uab.edu/weinstei/notes/dg.pdf

I.0 : Partial differentiation

(Lipschutz, Chapters 2, 6, 7)

This is an extremely brief review of the most basic facts that are covered in multivariable calculus courses.

The basic setting for multivariable calculus involves **Cartesian** or **Euclidean** n**-space**, which is denoted by $\mathbb{R}^n$. At first one simply takes $n = 2$ or 3 depending on whether one is interested in 2-dimensional or 3-dimensional problems, but much of the discussion also works for larger values of n. We shall view elements of these spaces as vectors, with addition and scalar multiplication done coordinatewise.

In order to do differential calculus for functions of two or more real variables easily, it is necessary to consider functions that are defined on **open sets**. One say of characterizing such a set is to say that $U \subset \mathbb{R}^n$ is open if and only if for each $\mathbf{p} = (p_1, ..., p_n) \in U$ there is an $\varepsilon > 0$ such that if $\mathbf{x} = (x_1, ..., x_n) \in U$ satisfies $|x_i - p_i| < \varepsilon$ for all i, then $\mathbf{x} \in U$. Alternatively, a set is open if and only if for each $\mathbf{p} \in U$ there is some $\delta > 0$ such that the set of all vectors $\mathbf{x}$ satisfying $|\mathbf{x} - \mathbf{p}| < \delta$ is contained in U (to see the equivalence of these for $n = 2$ or 3, consider squares inscribed in circles, squares circumscribed in circles, and similarly for cubes and spheres replacing squares and circles; illustrations and further discussion are in the files `neighborhoods.pdf` and `opensets.pdf`).

Continuous real valued functions on open sets are defined formally using the same sorts of $\varepsilon - \delta$ conditions that appear in single variable calculus; unless it is absolutely necessary, we shall try to treat such limits intuitively (for example, see the discussion in Section I.2). Vector valued functions are completely determined by the n scalar functions giving their coordinates, and a vector valued function is continuous if and only if all its scalar valued coordinate functions are continuous. As in single variable calculus, polynomials are always continuous, and standard constructions on continuous functions — for example, algebraic operations and forming composite functions – produce new continuous functions from old ones.

More generally, one can also define limits for functions of several variables either by means of the standard $\varepsilon - \delta$ condition; for functions of several variables, the appropriate condition for asking whether

$$\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x}) \quad = \quad b$$

is that the function f should be defined for all $\mathbf{x}$ sufficiently close to $\mathbf{a}$ with the possible exception of $\mathbf{x} = \mathbf{a}$. In other words, there is some $r > 0$ such that f is defined for all $\mathbf{x}$ satsisfying

$$0 \quad < \quad |\mathbf{x} - \mathbf{a}| \quad < r \ .$$

The definition of limit works equally well for vector and scalar valued functions, and the following basic result is often extremely useful when considering limits of vector valued functions.

VECTOR LIMIT FORMULA. *Let* $\mathbf{F}$ *be a vector valued function defined on a deleted neighborhood of* $\mathbf{a}$ *with values in* $\mathbb{R}^n$, *let* f_i *denote the* i^{th} *coordinate function of* $\mathbf{F}$, *and suppose that*

$$\lim_{\mathbf{x}\to\mathbf{a}} f_i(x) \quad = \quad b_i$$

holds for all i. *Let* $\mathbf{e}_i$ *denote the* i^{th} *unit vector in* $\mathbb{R}^n$, *whose* i^{th} *coordinate is equal to* 1 *and whose other coordinates are equal to zero. Then we have*

$$\lim_{\mathbf{x}\to\mathbf{a}} \mathbf{F}(x) \quad = \quad \sum_{i=1}^{n} b_i\,\mathbf{e}_i \ .\blacksquare$$

The previous statement about continuity of vector valued functions (continuous $\Longleftrightarrow$ all of the coordinate functions are continuous) is an immediate consequence of this formula.$\blacksquare$

Partial derivatives

Given a real valued function f defined on an open set U, its partial derivatives are formed as follows. For each index i between 1 and n, consider the functions obtained by holding all variables except the i^{th} variable constant, and take ordinary derivatives of such functions. The corresponding derivative is denoted by the standard notation

$$\frac{\partial f}{\partial x_i} \ .$$

The *gradient* of f is the vector ∇f whose i^{th} coordinate is equal to the i^{th} partial derivative.

One then has the following fundamentally important linear approximation result.

THEOREM. *Let* f *be a function defined on an open subset* $U \subset \mathbb{R}^n$, *and let* $\mathbf{x} \in U$. *Suppose also that* ∇f *is also continuous on* U. *Then there is a* $\delta > 0$ *and a function* θ *defined for* $|\mathbf{h}| < \delta$ *such that*

$$f(\mathbf{x} + \mathbf{h}) \quad = \quad f(\mathbf{x}) \ + \ \nabla f(\mathbf{x}) \cdot \mathbf{h} \ + \ |\mathbf{h}| \cdot \theta(\mathbf{h})$$

where $\lim_{|\mathbf{h}|\to 0} \ |\theta(\mathbf{h})| = 0$.$\blacksquare$

Derivations of this theorem are given in virtually every calculus book which devotes a chapter to partial differentiation. It is important to note that the existence of partial derivatives by itself is not even enough to ensure that a function is continuous (standard examples like

$$f(x, y) \quad = \quad \frac{x\,y}{x^2 + y^2}$$

for $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$ are also given in nearly all calculus books).

I.1 : Cross products

(Lipschutz, Chapter 1)

Courses in single variable or multivariable calculus usually define the cross product of two vectors and describe some of its basic properties. Since this construction will be particularly important to us and we shall use properties that are not always emphasized in calculus courses, we shall begin with a more detailed treatment of this construction.

Note on orthogonal vectors

One way of attempting to describe the dimension of a vector space is to suggest that the dimension represents the maximum number of mutually perpendicular directions. The following elementary result provides a formal justification for this idea.

PROPOSITION. *Let* $\mathbf{S} = \{\mathbf{a}_1, \cdots, \mathbf{a}_k\}$ *be a set of nonzero vectors that are mutually perpendicular. Then* $\mathbf{S}$ *is linearly independent.*

Proof. Suppose that we have an equation of the form

$$\sum_{i=1}^{n} c_i \mathbf{a}_i \quad = \quad \mathbf{0}$$

for some scalars c_i. If $1 \leq j \leq k$ we then have

$$0 \quad = \quad \mathbf{0} \cdot \mathbf{a}_j \quad = \quad \left(\sum_{i=1}^{n} c_i \mathbf{a}_i \right) \cdot \mathbf{a}_j \quad = \quad \sum_{i=1}^{n} (c_i \mathbf{a}_i \cdot \mathbf{a}_j)$$

and since the vectors in $\mathbf{S}$ are mutually perpendicular the latter reduces to $c_j|\mathbf{a}_j|^2$. Thus the original equation implies that $c_j|\mathbf{a}_j|^2 = 0$ for all j. Since each vector $\mathbf{a}j$ is nonzero it follows that $|\mathbf{a}_j|^2 > 0$ for all j which in turn implies $c_j = 0$ for all j. Therefore $\mathbf{S}$ is linearly independent.∎

Properties of cross products

Definition. If $\mathbf{a} = (a_1, a_2, a_3)$ and $\mathbf{b} = (b_1, b_2, b_3)$ are vectors in $\mathbb{R}^3$ then their cross product or vector product is defined to be

$$\mathbf{a} \times \mathbf{b} \quad = \quad \big(a_2b_3 - a_3b_2,\ a_3b_1 - a_1b_3,\ a_1b_2 - a_2b_1 \big) \ .$$

If we define unit vectors in the traditional way as $\mathbf{i} = (1, 0, 0)$, $\mathbf{j} = (0, 1, 0)$, and $\mathbf{k} = (0, 0, 1)$, then the right hand side may be written symbolically as a 3×3 deterinant:

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & a_2 & a_3 \end{vmatrix}$$

The following are immediate consequences of the definition:

(1) $\mathbf{a} \times \mathbf{b} \quad = \quad -\mathbf{b} \times \mathbf{a}$

(2) $(c\,\mathbf{a}) \times \mathbf{b} \quad = \quad c\,(\mathbf{a} \times \mathbf{b})$

(3) $\mathbf{a} \times (\mathbf{b} + \mathbf{c}) \quad = \quad (\mathbf{a} \times \mathbf{b}) \ + \ (\mathbf{a} \times \mathbf{c})$

Other properties follow directly. For example, by (1) we have that $\mathbf{a} \times \mathbf{a} = -\mathbf{a} \times \mathbf{a}$, so that $2\,\mathbf{a} \times \mathbf{a} = \mathbf{0}$, which means that $\mathbf{a} \times \mathbf{a} = \mathbf{0}$. Also, if $\mathbf{c} = (c_1, c_2, c_3)$ then the triple product

$$[\mathbf{c}, \mathbf{a}, \mathbf{b}] \quad = \quad \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})$$

is simply the determinant of the 3×3 matrix whose rows are $\mathbf{c}$, $\mathbf{a}$, $\mathbf{b}$ in that order, and therefore we know that

the cross product $\mathbf{a} \times \mathbf{b}$ *is perpendicular to both* $\mathbf{a}$ *and* $\mathbf{b}$.■

The direction of the cross product is determined by this perpendicularity property and a rule for determining which way the cross product points along this perpendicular, which is stated in various forms and known as the **right hand rule**. Two examples to illustrate the right hand rule appear on pages 11 and 12 of the following document:

`http://math.ucr.edu/~res/math133/geometrynotes1.pdf`

The basic properties of determinants also yield the following additional identity involving dot and cross products:

$$[\mathbf{c}, \mathbf{a}, \mathbf{b}] \quad = \quad [\mathbf{a}, \mathbf{b}, \mathbf{c}]$$

This follows because a determinant changes sign if two rows are switched, for the latter implies

$$[\mathbf{c}, \mathbf{a}, \mathbf{b}] \quad = \quad -[\mathbf{a}, \mathbf{c}, \mathbf{b}] \quad = \quad [\mathbf{a}, \mathbf{b}, \mathbf{c}]\ .■$$

The following property of cross products plays an extremely important role in this course.

PROPOSITION. *If* $\mathbf{a}$ *and* $\mathbf{b}$ *are linearly independent, then* $\mathbf{a}$, $\mathbf{b}$ *and* $\mathbf{a} \times \mathbf{b}$ *form a basis for* $\mathbb{R}^3$.

Proof. First of all, we claim that if $\mathbf{a}$ and $\mathbf{b}$ are linearly independent, then $\mathbf{a} \times \mathbf{b} \neq \mathbf{0}$. To see this we begin by writing out $|\mathbf{a} \times \mathbf{b}|^2$ explicitly:

$$|\mathbf{a} \times \mathbf{b}|^2 \quad = \quad (a_2 b_3 - a_3 b_2)^2 + (a_3 b_1 - a_1 b_3)^2 + (a_1 b_2 - a_2 b_1)^2$$

Direct computation shows that the latter is equal to

$$(a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) \ - \ (a_1 b_1 + a_2 b_2 + a_3 b_3)^2 \quad = \quad |\mathbf{a}|^2\,|\mathbf{b}|^2 \ - \ (\mathbf{a} \cdot \mathbf{b})^2$$

In particular, if $\mathbf{a}$ and $\mathbf{b}$ are both nonzero then

$$|\mathbf{a} \times \mathbf{b}| \quad = \quad |\mathbf{a}|\,|\mathbf{b}|\,|\sin\theta|$$

where θ is the angle between $\mathbf{a}$ and $\mathbf{b}$. Since the sine of this angle is zero if and only if the vectors are linearly dependent, it follows that $\mathbf{a} \times \mathbf{b} \neq \mathbf{0}$ if $\mathbf{a}$ and $\mathbf{b}$ are linearly independent.

Suppose now that we have an equation of the form

$$x\,\mathbf{a} + y\,\mathbf{b} + z(\mathbf{a}\times\mathbf{b}) \quad = \quad \mathbf{0}$$

for suitable scalars x, y, z. Taking dot products with $\mathbf{a}\times\mathbf{b}$ yields the equation $z|\mathbf{a}\times\mathbf{b}|^2 = 0$, which by the previous paragraph implies that $z = 0$. One can now use the linear independence of $\mathbf{a}$ and $\mathbf{b}$ to conclude that x and y must also be zero. Therefore the three vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{a}\times\mathbf{b}$ are linearly independent, and consequently they must form a basis for $\mathbb{R}^3$.■

APPLICATION. Later in these notes we shall need the following result:

RECOGNITION FORMULA. *If* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$ *are perpendicular unit vectors and* $\mathbf{c} = \mathbf{a}\times\mathbf{b}$, *then the triple product* $[\mathbf{a}, \mathbf{b}, \mathbf{c}]$ *is equal to* 1.

Derivation. By the length formula for a cross product and the perpendicularity assumption, we know that $|\mathbf{c}| = |\mathbf{a}|\cdot|\mathbf{b}| = 1\cdot 1 = 1$. But we also have

$$1 \quad = \quad |\mathbf{c}|^2 \quad = \quad \mathbf{c}\cdot(\mathbf{a}\times\mathbf{b}) \quad = \quad [\mathbf{c}, \mathbf{a}, \mathbf{b}] \quad = \quad [\mathbf{a}, \mathbf{b}, \mathbf{c}]$$

which is the equation that we want.■

Cross products of three vectors

In may situations it is useful to have formulas for more complicated expressions involving cross products. For example, we have the following identity for computing threefold cross products.

"BAC—CAB" RULE. $\mathbf{a}\times\big(\mathbf{b}\times\mathbf{c}\big) = \mathbf{b}(\mathbf{a}\cdot\mathbf{c}) - \mathbf{c}(\mathbf{a}\cdot\mathbf{b})$, *or in more standard format the left hand side is equal to* $(\mathbf{a}\cdot\mathbf{c})\,\mathbf{b} - (\mathbf{a}\cdot\mathbf{b})\,\mathbf{c}$.

Derivation. Suppose first that $\mathbf{b}$ and $\mathbf{c}$ are linearly dependent. Then their cross product is zero, and one is a scalar multiple of the other. If $\mathbf{b} = x\,\mathbf{c}$, then it is an elementary exercise to verify that the right hand side of the desired identity is zero, and we already know the same is true of the left hand side. If on the other hand $\mathbf{c} = y\,\mathbf{b}$, then once again one finds that both sides of the desired identity are zero.

Now suppose that $\mathbf{b}$ and $\mathbf{c}$ are linearly independent, so that $\mathbf{b}\times\mathbf{c} \neq \mathbf{0}$. Note that a vector is perpendicular to $\mathbf{b}\times\mathbf{c}$ if and only if it is a linear combination of $\mathbf{b}$ and $\mathbf{c}$. The ($\Longleftarrow$) implication follows from the perpendicularity of $\mathbf{b}$ and $\mathbf{c}$ to their cross product and the distributivity of the dot product, while the reverse implication follows because every vector is a linear combination

$$x\,\mathbf{b} + y\,\mathbf{c} + z\,(\mathbf{b}\times\mathbf{c})$$

and this linear combinationn is perpendicular to the cross product if and only if $z = 0$; *i.e.*, if and only if the vector is a linear combination of $\mathbf{b}$ and $\mathbf{c}$.

Before studying the general case, we shall first consider the special cases $\mathbf{b}\times(\mathbf{b}\times\mathbf{c})$ and $\mathbf{c}\times\big(\mathbf{b}\times\mathbf{c}\big)$. Since $\mathbf{b}\times\big(\mathbf{b}\times\mathbf{c}\big)$ is perpendicular to $\mathbf{b}\times\mathbf{c}$ we may write it in the form

$$\mathbf{b}\times\big(\mathbf{b}\times\mathbf{c}\big) \quad = \quad u\,\mathbf{b} + v\,\mathbf{c}$$

for suitable scalars u and v. If we take dot products with $\mathbf{b}$ and $\mathbf{c}$ we obtain the following equations:

$$\mathbf{0} \quad = \quad [\mathbf{b}, \mathbf{b}, \mathbf{b}\times\mathbf{c}] \quad = \quad \Big(\mathbf{b}\cdot\big(\mathbf{b}\times(\mathbf{b}\times\mathbf{c})\,\big)\Big) \quad = \quad \mathbf{b}\cdot\big(u\,\mathbf{b} + v\,\mathbf{c}\big) \quad = \quad u\,(\mathbf{b}\cdot\mathbf{b}) + v\,(\mathbf{b}\cdot\mathbf{c})$$

$$-|\mathbf{b}\times\mathbf{c}|^2 \;=\; -[(\mathbf{b}\times\mathbf{c}),\mathbf{b},\mathbf{c}] \;=\; [\mathbf{b},(\mathbf{b}\times\mathbf{c}),\mathbf{c}] \;=\; [\mathbf{c},\mathbf{b},(\mathbf{b}\times\mathbf{c})] \;=$$
$$\Big(\mathbf{c}\cdot\big(\mathbf{b}\times(\mathbf{b}\times\mathbf{c})\;\big)\Big) \;=\; \mathbf{c}\cdot\big(u\,\mathbf{b}+v\,\mathbf{c}\big) \;=\; u\,(\mathbf{b}\cdot\mathbf{c})+v\,(\mathbf{c}\cdot\mathbf{c})$$

If we solve these equations for u and v we find that $u=\mathbf{b}\cdot\mathbf{c}$ and $v=-\mathbf{b}\cdot\mathbf{b}$. Therefore we have

$$\mathbf{b}\times\big(\mathbf{b}\times\mathbf{c}\big) \;=\; (\mathbf{b}\cdot\mathbf{c})\,\mathbf{b}-(\mathbf{b}\cdot\mathbf{b})\,\mathbf{c}\;.$$

Similarly, we also have

$$\mathbf{c}\times\big(\mathbf{b}\times\mathbf{c}\big) \;=\; (\mathbf{c}\cdot\mathbf{c})\,\mathbf{b}-(\mathbf{b}\cdot\mathbf{c})\,\mathbf{c}\;.$$

If we now write $\mathbf{a}=p\,\mathbf{b}+q\,\mathbf{c}+r\big(\mathbf{b}\times\mathbf{c}\big)$ we have

$$\mathbf{a}\times\big(\mathbf{b}\times\mathbf{c}\big) \;=\; p\,\mathbf{b}\times\big(\mathbf{b}\times\mathbf{c}\big) \;+\; q\,\mathbf{c}\times\big(\mathbf{b}\times\mathbf{c}\big) \;=$$
$$\Big(p\,(\mathbf{b}\cdot\mathbf{c})+q\,(\mathbf{c}\cdot\mathbf{c})\Big)\,\mathbf{b} \;-\; \Big(p\,(\mathbf{b}\cdot\mathbf{b})+q\,(\mathbf{b}\cdot\mathbf{c})\Big)\,\mathbf{c}\;.$$

Since $\mathbf{b}$ and $\mathbf{c}$ are perpendicular to their cross product, we must have $(\mathbf{a}\cdot\mathbf{c})=p\,(\mathbf{b}\cdot\mathbf{c})+q\,(\mathbf{c}\cdot\mathbf{c})$ and $(\mathbf{a}\cdot\mathbf{b})=p\,(\mathbf{b}\cdot\mathbf{b})+q\,(\mathbf{b}\cdot\mathbf{c})$, so that the previously obtained expression for $\mathbf{a}\times(\mathbf{b}\times\mathbf{c})$ is equal to $(\mathbf{a}\cdot\mathbf{c})\,\mathbf{b}-(\mathbf{a}\cdot\mathbf{b})\,\mathbf{c}$.■

The formula for $\mathbf{a}\times\big(\mathbf{b}\times\mathbf{c}\big)$ yields numerous other identities. Here is one that will be particularly useful in this course.

PROPOSITION. *If* $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ *and* $\mathbf{d}$ *are arbitrary vectors in* $\mathbb{R}^3$ *then we have the following identity:*

$$\big(\mathbf{a}\times\mathbf{b}\big)\cdot\big(\mathbf{c}\times\mathbf{d}\big) \;=\; (\mathbf{a}\cdot\mathbf{c})(\mathbf{b}\cdot\mathbf{d}) \;-\; (\mathbf{a}\cdot\mathbf{d})(\mathbf{b}\cdot\mathbf{c})$$

Proof. By definition, the expression on the left hand side of the display is equal to the triple product $[(\mathbf{a}\times\mathbf{b}),\mathbf{c},\mathbf{d}]$. As noted above, the properties of determinants imply that the latter is equal to $[\mathbf{d},(\mathbf{a}\times\mathbf{b}),\mathbf{c}]$, which in turn is equal to

$$\mathbf{d}\cdot\big(\mathbf{a}\times\big(\mathbf{b}\times\mathbf{c}\big)\big) \;=\; \mathbf{d}\cdot\big(\,(\mathbf{a}\cdot\mathbf{c})\mathbf{b}-(\mathbf{a}\cdot\mathbf{b})\mathbf{c}\,\big)$$

and if we expand the final term we obtain the expression $(\mathbf{a}\cdot\mathbf{c})(\mathbf{b}\cdot\mathbf{d})-(\mathbf{a}\cdot\mathbf{d})(\mathbf{b}\cdot\mathbf{c})$.■

Cross products and higher dimensions

Given the relative ease in defining generalizations of the inner (or dot) product and the usefulness of the 3-dimensional cross product in mathematics and physics, it is natural to ask whether there are also generalizations of the cross product. However, it is rarely possible to define good generalizations of the cross product that satisfy most of the latter's good properties. Partial but significantly more complicated generalizations can be constructed using relatively sophisticated techniques (for example, from tensor algebra or Lie algebras), but such material goes far beyond the scope of this course. Here are two online references containing further information:

`http://www.math.niu.edu/~rusin/known-math/95/prods`

`http://www.math.niu.edu/~rusin/known-math/96/octonionic`

We shall not use the material in these reference subsequently.

Although one does not have good theories of cross products in higher dimensions, there is a framework for generalizing many important features of this construction to higher dimensions. This is implicit in the theory of **differential forms**; a discussion of the 2- and 3-dimensional cases appears in Section II.1 of these notes.

Appendix: The distance between two skew lines

To illustrate the uses of calculus and linear algebra to work geometric problems, we shall prove a basic result on **skew lines**; *i.e.*, lines which have no points in common but are not parallel (hence they cannot be coplanar). It follows that two lines are skew lines if and only if there are two points p, q on one and two points p', q' on the other such that the set $\{p, q, p', q'\}$ is not coplanar. Two drawings of skew lines are in the course directory file `skew-lines.pdf`.

THEOREM. *Let L and M be two skew lines in $\mathbb{R}^3$, and for $\mathbf{x} \in L$ and $\mathbf{y} \in M$ let $d(\mathbf{x}, \mathbf{y})$ denote the distance between $\mathbf{x}$ and $\mathbf{y}$. Then the function $d(\mathbf{x}, \mathbf{y})$ takes a positive minimum value, and if $\mathbf{x}_m$ and $\mathbf{y}_m$ are points where $d(\mathbf{x}, \mathbf{y})$ is minimized, then the line joining $\mathbf{x}_m$ and $\mathbf{y}_m$ is perpendicular to both L and M.*

In classical Euclidean geometry this is usually stated in the form, "The shortest distance between two skew lines is along their common perpendicular." Predictably, it is possible to prove this result using the methods of classical synthetic geometry, and nearly all the textbooks on solid geometry from the first six decades of the 20^{th} century contain proofs of this result.

Proof. There are three main parts to the argument:

(1) Proving that the distance function has an absolute minimum; under the hypotheses, we know that this minimum distance must be positive.

(2) Showing that the the minimum value is realized by points $\mathbf{x}_m$ and $\mathbf{y}_m$ such that the line $\mathbf{x}_m \, \mathbf{y}_m$ is perpendicular to both L and M.

(3) Deriving an algebraic formula for the minimum distance; one version of this formula is given in Problem 8 on page 15 of DO CARMO.

FIRST STEP. We begin by translating the problem into a question about vectors. Suppose that the skew lines have parametric equations of the form

$$\mathbf{p}_0 \; + \; t\,\mathbf{u} \quad , \quad \mathbf{p}_1 \; + \; s\,\mathbf{v}$$

where $\mathbf{u}$ and $\mathbf{v}$ are nonzero and in fact must be *linearly independent*; for if $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent then the two lines described above are identical or parallel. In effect the problem is to show that the function $f(s,t) \;=\; |\,\mathbf{r}(s,t)\,|^2$, where

$$\mathbf{r}(s,t) \;=\; \big(\mathbf{p}_0 \; + \; t\,\mathbf{u}\big) \; - \; \big(\mathbf{p}_1 \; + \; s\,\mathbf{v}\big)$$

has a minimum value and to find that value.

As noted above, we shall begin by provig that there is a minimum value. If we write out the conditions for a point to satisfy $\nabla f(s,t) = \mathbf{0}$ we obtain the following system of linear equations, where A and B are some constants.

$$t\,\langle \mathbf{u}, \mathbf{u}\rangle \; - \; s\,\langle \mathbf{u}, \mathbf{v}\rangle \;=\; A$$

$$t\,\langle \mathbf{u}, \mathbf{v}\rangle \; - \; s\,\langle \mathbf{v}, \mathbf{v}\rangle \;=\; B$$

These equations have a unique solution because the determinant

$$\begin{vmatrix} \langle \mathbf{u}, \mathbf{u}\rangle & \langle \mathbf{u}, \mathbf{v}\rangle \\ \langle \mathbf{u}, \mathbf{v}\rangle & \langle \mathbf{v}, \mathbf{v}\rangle \end{vmatrix}$$

is nonzero by the Schwarz inequality and the linear indepdendence of $\mathbf{u}$ and $\mathbf{v}$. Let $R > 0$ be so large that the solution $(s^*,\, t^*)$ lies inside the circle $s^2+t^2 = R^2$. Then on the set $s^2+t^2 = R^2$ either the minimum value occurs at the unique critical point or else it occurs on the boundary circle. Let D be the value of the function at the critical point, so that $D \geq 0$. If D is not a minimum value for $f(s,t)$ then for every $Q > R$ there is a point on the circle $s^2 + t^2 = Q^2$ for which the value of the function is less than D. We claim this is impossible, and it will follow that D must be the minimum value of the function.

Consider the values of the function f on the circle of radius ρ; these are given by

$$|\,\mathbf{r}\,(\,\rho\,\cos\theta,\ \rho\,\sin\theta\,)\,|^2$$

and if we write everything out explicitly we obtain the following expression for this function, in which $\mathbf{q}$ is the vector $\mathbf{p}_0 - \mathbf{p}_1$:

$$\rho^2\,|\cos\theta\,\mathbf{u} - \sin\theta\,\mathbf{v}|^2 \ + \ 2\,\rho\,\langle\cos\theta\mathbf{u} - \sin\theta\,\mathbf{v},\, \mathbf{q}\rangle \ + \ |\mathbf{q}|^2$$

Let m denote the minimum value of $|\cos\theta\,\mathbf{u} - \sin\theta\,\mathbf{v}|$ for $\theta \in [0,\, 2\pi]$ and let M denote the maximum value. Since $\mathbf{u}$ and $\mathbf{v}$ are linearly independent, the displayed expression is always positive and therefore m must be positive. We claim that the minimum value of $f(s,t)$ on the circle $s^2+t^2 = \rho^2$ is greater than or equal to the following expression:

$$\rho^2\,m^2 - 2\,\rho\,M\,|\mathbf{q}| + |\mathbf{q}|^2$$

This follows immediately from the inequalities

$$\rho^2\,|\cos\theta\,\mathbf{u} - \sin\theta\,\mathbf{v}|^2 \quad \geq \quad \rho^2\,m^2$$

$$2\,\rho\,\langle\cos\theta\mathbf{u} - \sin\theta\,\mathbf{v},\, \mathbf{q}\rangle \quad \geq \quad -\,2\rho\,|\cos\theta\,\mathbf{u} - \sin\theta\,\mathbf{v}|\cdot|\mathbf{q}| \quad \geq 2\,\rho\,M\,|\mathbf{q}|$$

where the first inequality in the second line comes from the Schwarz inequality.

Since

$$\lim_{\rho\to\infty}\ \rho^2\,m^2 - 2\,\rho\,M\,|\mathbf{q}| + |\mathbf{q}|^2 \quad = \quad +\infty$$

it follows that all sufficiently large ρ the minimum value of $f(s,t)$ on the circle $s^2+t^2 = \rho^2$ is strictly greater than D, and therefore D must be the absolute minimum for f on the set $s^2 + t^2 \leq \rho^2$ for all sufficiently large ρ. But this means that D must be the absolute minimum for the function over all possible values of s and t.

SECOND STEP. In order to determine where the minimum value is attained, one must set the partial derivatives of f with respect to a and t both equal to zero. If we do this we obtain the following equations:

$$0 \quad = \quad 2\,\mathbf{r}(s,t)\cdot(-\mathbf{v})$$

$$0 \quad = \quad 2\,\mathbf{r}(s,t)\cdot(\ \mathbf{u})$$

Since $\mathbf{u}$ and $\mathbf{v}$ are linearly independent, this minimum occurs when $\mathbf{r}(s,t)$ a scalar multiple of $\mathbf{u}\times\mathbf{v}$.

Suppose that the minimum distance between the lines is attained at parameter values (s_0, t_0). If $\mathbf{x}$ and $\mathbf{y}$ are the points on the lines where this minimum value is realized, then by construction we know that $\mathbf{r}(s_0, t_0) = \mathbf{x} - \mathbf{y}$, and since the left hand side is a multiple of $\mathbf{u}\times\mathbf{v}$ it follows that the line joining $\mathbf{x}$ and $\mathbf{y}$ is perpendicular to both L and M.

THIRD STEP. As above, Suppose that the minimum distance between the lines is attained at parameter values (s_0, t_0). Then as before we have $\mathbf{r}(s_0, t_0) = k\,\mathbf{u} \times \mathbf{v}$ for some scalar k, and it follows immediately that the minimum distance d satisfies

$$d \quad = \quad \frac{\left|\ \left[\mathbf{u}, \mathbf{v}, \mathbf{r}(s_0, t_0)\right]\ \right|}{|\mathbf{u} \times \mathbf{v}|}$$

where $[\mathbf{a}, \mathbf{b}, \mathbf{c}]$ refers to the usual triple product of vectors having the form $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$. The exercise in do Carmo claims that a similar formula holds with $\mathbf{r}(0,0) = \mathbf{p}_0 = \mathbf{p}_1$ replacing $\mathbf{r}(s_0, t_0)$. This is true because

$$\mathbf{r}(0,0) = \mathbf{r}(s_0, t_0) \ + \ t_0\,\mathbf{u} \ - \ s_0\,\mathbf{v_0}$$

which implies that the triple products $\left[\mathbf{u}, \mathbf{v}, \mathbf{r}(s_0, t_0)\right]$ and $\left[\mathbf{u}, \mathbf{v}, \mathbf{r}(0,0)\right]$ are equal. This is the formula in DO CARMO.

A simpler example of a geometric proof using vectors is in the file `vector-proofs.pdf`.

I.2 : Parametrized curves

(Lipschutz, Chapter 3)

There is a great deal of overlap between the contents of this section and certain standard topics in calculus courses. One major difference in this course is the need to work more systematically with some fundamental but relatively complex theoretical points in calculus that can (and in most cases probably should) be overlooked when working most ordinary and multivariable calculus problems. In particular this applies to the definitions of limits and continuity, and accordingly we shall begin with some comments on this background material.

Useful facts about limits

In ordinary and multivariable calculus courses it is generally possible to get by with only a vague understanding of the concept of limit, but in this course a somewhat better understanding is necessary. In particular, the following consequences of the definition arise repeatedly.

FACT I. *Let f be a function defined at all points of the interval $(a-h, a+h)$ for some $h > 0$ except possibly at a, and suppose that*

$$\lim_{x \to a} f(x) = b > 0 \ .$$

Then there is a $\delta > 0$ such that $\delta < h$ and $f(x) > 0$ provided $x \in (a-\delta, a+\delta)$ and $x \neq a$.

FACT II. *In the situation described above, if the limit exists but is* **negative***, then there is a $\delta > 0$ such that $\delta < h$ and $f(x) > 0$ provided $x \in (a-\delta, a+\delta)$ and $x \neq a$.*

FACT III. *Each of the preceding statements remains true if 0 is replaced by an arbitrary real number.*

Derivation(s). We shall only do the first one; the other two proceed along similar lines. By assumption b is a positive real number. Therefore the definition of limit implies there is some $\delta > 0$ such that $|f(x) - b| < b$ provided provided $x \in (a - \delta, a + \delta)$ and $x \neq a$. It then follows that

$$f(x) \quad = \quad b \; + \; \big(f(x) \; - \; b \big) \quad \geq \quad b \; - \; |f(x) - b| \quad > \quad b \; - \; b \quad = \quad 0$$

which is what we wanted to show.■

We shall also need the following statement about infinite limits:

FACT IV. *Let f be a continuous function defined on some open interval containing 0 such that f is strictly increasing and $f(0) = 0$. Then for each positive constant C there is a positive real number h sufficiently close to zero such that $x \in (0,\, h) \implies 1/f(x) > C$ and $x \in (-h,\, 0) \implies 1/f(x) < -C$.*

Proof. Let ε be the positive number $1/C$; by continuity we know that $|f(x)| < \varepsilon$ if $x \in (-h,\, h)$ for a suitably small $h > 0$. Therefore $x \in (0,\, h) \implies 0 < f(x) < \varepsilon$ and $x \in (-h,\, 0) \implies -\varepsilon < f(x) < 0$. The desired inequalities follow by taking reciprocals in each case.■

What is a curve?

There are two different but related ways to think about curves in the plane or 3-dimensional space. One can view a curve simply as a set of points, or one can view a curve more dynamically as a description of the position of a moving object at a given time. In calculus courses one generally adopts the second approach to define curves in terms of parametric equations; from this viewpoint one retrieves the description of curves as sets of points by taking the set of all points traced out by the moving object, where the independent "time" variable lies in some interval J. For example, the line in $\mathbb{R}^2$ defined by the equation $y = m\,x$ is the set of points traced out by the parametrized curve defined by $x(t) = t$ and $y(t) = m\,t$. Similarly, the unit circle defined by the equation $x^2 + y^2 = 1$ is the set of points traced out by the parametrized curve $x(t) = \cos t$, $y(t) = \sin t$. The set of all points expressible as $\mathbf{x}(t)$ for some $t \in J$ will be called the *image* of the parametrized curve (since it represents all point traced out by the curve this set is sometimes called the *trace* of the curve, but we shall not use this term in order to avoid confusion with the entirely different notion of the trace of a matrix). We shall follow the standard approach of calculus books here unless stated otherwise.

A parametrized curve in the plane or 3-dimensional space may be viewed as a vector-valued function γ or $\mathbf{x}$ defined on some interval of the real line and taking values in $V = \mathbb{R}^2$ or $\mathbb{R}^3$. In this course we usually want our curves to be continuous; this is equivalent to saying that each of the coordinate functions is continuous. Given that this is a course in *differential* geometry it should not be surprising that we also want our curves to have some decent differentiability properties. If $\mathbf{x}$ is the vector function defining our curve and its coordinates are given by x_i, where i runs between 1 and 2 or 1 and 3 depending upon the dimension of V, then the derivative of $\mathbf{x}$ at a point t is defined using the coordinate functions:

$$\mathbf{x}'(t) = \big(\, x_1'(t),\, x_2'(t),\, x_3'(t) \,\big)$$

Strictly speaking this is the definition in the 3-dimensional case, but the adaptation to the 2-dimensional case is immediate — one can just suppress the third coordinate or view $\mathbb{R}^2$ as the subset of $\mathbb{R}^3$ consisting of all points whose third coordinate is zero.

Definition. A curve $\mathbf{x}$ defined on an interval J and taking values in $V = \mathbb{R}^2$ or $\mathbb{R}^3$ is *differentiable* if $\mathbf{x}'(t)$ exists for all $t \in J$. The curve is said to be *smooth* if $\mathbf{x}'$ is continuous, and it is said to be a *regular smooth curve* if it is smooth and $\mathbf{x}'(t)$ is nonzero for all $t \in J$. The curve will be said to be *smooth of class* $\mathcal{C}^r$ for some integer $r \geq 1$ if $\mathbf{x}$ has an r^{th} order continuous derivative, and the curve will be said to be smooth of class $\mathcal{C}^\infty$ if it is infinitely differentiable (equivalently, $\mathcal{C}^r$ for all finite r).

The crucial property of regular smooth curves is that they have well defined tangent lines:

Definition. Let $\mathbf{x}$ be a regular smooth curve and let a be a point in the domain J of $\mathbf{x}$. The *tangent line* to $\mathbf{x}$ at the parameter value $t = a$ is the unique line passing through $\mathbf{x}(a)$ and $\mathbf{x}(a) + \mathbf{x}'(a)$. There is a natural associated parametrization of this line given by

$$T(u) = \mathbf{x}(a) \ + \ u\,\mathbf{x}'(a) \ .$$

One expects the tangent line to be the "best possible" linear approximation to a smooth curve. The following result confirms this:

PROPOSITION. *In the notation above, if $u \neq 0$ is small and $a + u \in J$ then we have*

$$\mathbf{x}(a+u) \quad = \quad \mathbf{x}(a) \ + \ u\,\mathbf{x}'(a) \ + \ u\,\Theta(u)$$

where $\lim_{u\to 0} \Theta(u) = \mathbf{0}$. Furthermore, if $\mathbf{p}$ is any vector such that

$$\mathbf{x}(a+u) \quad = \quad \mathbf{x}(u) \ + \ u\,\mathbf{p} \ + \ u\,\mathbf{W}(u)$$

where $\lim_{u\to 0} \mathbf{W}(u) = \mathbf{0}$, then $\mathbf{p} = \mathbf{x}'(a)$.

Proof. Given a vector $\mathbf{a}$ we shall denote its i^{th} coordinate by a_i.

Certainly there is no problem writing $\mathbf{x}(a+u)$ in the form $\mathbf{x}(u) + u\,\mathbf{x}'(a) + u\,\Theta(u)$ for some vector valued function Θ; the substance of the first part of the proposition is that this function goes to zero as $u \to 0$. Limit identities for vector valued functions are equivalent to scalar limit identities for every coordinate function of the vectors, so the proof of the first part of the proposition reduces to checking that the coordinates θ_i of Θ satisfy $\lim_{u\to 0} \theta_i(u) = 0$ for all i. However, by construction we have

$$\theta_i(u) \quad = \quad \frac{x_i(a+u) - x_i(a)}{u} \ - \ x_i'(a)$$

and since $\mathbf{x}$ is differentiable at a the limit of the right hand side of this equation is zero. Therefore we have where $\lim_{u\to 0} \Theta(u) = \mathbf{0}$.

Suppose now that the second equation in the statement of the proposition is valid. As in the previous paragraph we have

$$w_i(u) \quad = \quad \frac{x_i(a+u) - x_i(a)}{u} \ - \ p_i(a)$$

but this time we know that $\lim_{u\to 0} w_i(u) = 0$ for all i. The only way these equations can hold is if $p_i(a) = x_i'(a)$ for all i.■

Piecewise smooth curves

There are many important geometrical curves that that are not smooth but can be decomposed into smooth pieces. One of the simplest examples is the boundary of the square parametrized in a counterclockwise sense. Specifically, take $\mathbf{x}$ to be defined on the interval $[0,4]$ by the following rules:

$$\begin{array}{llcll} (a) & \mathbf{x}(t) & = & (t,0) & \text{for } t \in [0,1] \\ (b) & \mathbf{x}(t) & = & (1,t-1) & \text{for } t \in [1,2] \\ (c) & \mathbf{x}(t) & = & (2-t,1) & \text{for } t \in [2,3] \\ (d) & \mathbf{x}(t) & = & (0,1-t) & \text{for } t \in [3,4] \end{array}$$

The formulas for (a) and (b) agree when $t = 1$, and likewise the formulas for (b) and (c) agree when $t = 2$, and finally the formulas for (c) and (d) agree when $t = 3$; therefore these formulas define a continuous curve. On each of the intervals $[n, n+1]$ for $n = 0, 1, 2, 3$ the curve is a regular smooth curve, but of course the tangent vectors coming from the left and the right at these values are perpendicular to each other. Clearly there are many other examples of this sort, and they include all broken line curves. The following definition includes both these types of curves and regular smooth curves as special cases:

Definition. A continuous curve $\mathbf{x}$ defined on an interval $[a,b]$ is said to be a *regular piecewise smooth curve* if there is a partition of the interval given by points

$$a \;=\; p_0 \;<\; p_1 \;\cdots\; <\; p_{n-1} \;< p_n \;= b$$

such that for each i the restriction $\mathbf{x}[i]$ of $\mathbf{x}$ to the subinterval $[p_{i-1}, p_i]$ is a regular smooth curve.

For the boundary of the square parametrized in the counterclockwise sense, the partition is given by

$$0 \;<\; 1 \;<\; 2 \;<\; 3 \;<\; 4\,.$$

Calculus texts give many further examples of such curves, and the references cited at the beginning of this unit also contain a wide assortment of examples. One important thing to note is that at each of the partition points p_i one has a left hand tangent vector $\mathbf{x}'(p_i-)$ obtained from $\mathbf{x}[i]$ and a right hand tangent vector $\mathbf{x}'(p_i+)$ obtained from $\mathbf{x}[i+1]$, but these two vectors are not necessarily the same. In particular, they do not coincide at the partition points $1, 2, 3$ for the parametrized boundary curve for the square that was described above.

Taylor's Formula for vector valued functions

We shall need a vector analog of the usual Taylor's Theorem for polynomial approximations of real valued functions on an interval.

VECTOR VALUED TAYLOR'S THEOREM. *Let $\mathbf{g}$ be a vector valued function defined on an interval $(a-r, a+r)$ that has continuous derivatives of all orders less than or equal to $n+1$ on that interval. Then for $|h| < r$ we have*

$$\mathbf{g}(a+h) \;=\; \mathbf{g}(a) \;+\; \sum_{k=1}^{n} \frac{h^k}{k!}\,\mathbf{g}^{(k)}(a) \;+\; \int_a^{a+h} \frac{(a+h-t)^n}{n!}\,\mathbf{g}^{(n+1)}(t)\,dt$$

where $\mathbf{g}^{(k)}$ as usual denotes the k^{th} derivative of $\mathbf{g}$.

Proof. Let $R_n(h)$ be the integral in the displayed equation. Then integration by parts implies that

$$R_{n-1}(h) \quad = \quad \frac{h^n}{n!}\,\mathbf{g}^{(n)}(a) \ + \ R_n(h)$$

and the Fundamental Theorem of Calculus implies that

$$\mathbf{g}(a+h) \quad = \quad \mathbf{g}(a) \ + \ R_1(h) \ .$$

Therefore if we set $R_0 = 0$ we have

$$\mathbf{g}(a+h) \quad = \quad \mathbf{g}(a) \ + \ \sum_{k=1}^{n} \big(R_k(h) - R_{k-1}(h)\big) \ + \ R_n(h)$$

and if we use the formulas above to substitute for the terms $\mathrm{R}_k(h) - R_{k-1}(h)$ and $R_n(h)$ we obtain the formula displayed above.■

The following consequence of Taylor's Theorem will be particularly useful:

COROLLARY. *Given* $\mathbf{g}$ *and the other notation as above, let* $P_n(h)$ *be the sum of*

$$\mathbf{g}(a) \ + \ \sum_{k=1}^{n} \frac{h^k}{k!}\,\mathbf{g}^{(k)}(a) \ .$$

Then given $r_0 < r$ *and* $|h| < r_0 < r$ *we have* $|\mathbf{g}(a+h) - P_n(h)| \leq C\,|h|^{n+1}$*, for some positive constant* C*.*

If we think of $\mathbf{g}$ as defining a parametrized curve and take $n = 1$, so that $\mathbf{g}''$ is continuous, then the corollary implies that the distance between the point $\mathbf{g}(a+t)$ on the curve and the corresponding point on the best linear approximation

$$\mathbf{g}(a) \ + \ t\,\mathbf{g}'(a)$$

is bounded from above by $C\,t^2$ for some constant C.

Proof of the Corollary. The length of the difference vector in the previous sentence is given by

$$|R_n(h)| \quad = \quad \left| \int_a^{a+h} \frac{(a+h-t)^n}{n!}\,\mathbf{g}^{(n+1)}(t)\,dt \right| \quad \leq$$

$$\mathrm{sign}(h) \ \cdot \ \int_a^{a+h} \left| \frac{(a+h-t)^n}{n!}\,\mathbf{g}^{(n+1)}(t) \right| \,dt \quad \leq$$

$$\Big(\max_{|t-a|\leq r_0} |\mathbf{g}^{(n+1)}(t)|\Big) \cdot \int_0^{|h|} \frac{u^n}{n!}\,du \quad \leq \quad M\,\frac{|h|^{n+1}}{(n+1)!}$$

where M is a positive constant at least as large as the maximum value of $|\mathbf{g}^{(n+1)}(t)|$ for $|t-a| < r_0$.■

NOTE. For each positive integer n there are functions which have continuous derivatives of order $\leq n$ but no globally defined derivative of order $n+1$. Examples are discussed at the end of this section.

Algebraic and transcendental curves

Frequently curves are defined by means of an equation of the form $F(x,y) = 0$, where F is a function of two variables with continuous partial derivatives. Normally one makes the following additional assumption:

At each point where $F(x,y) = 0$ *we have* $\nabla F(x,y) \neq \mathbf{0}$.

If this condition is met at (a, b) such that $F(a, b)$, then the Implicit Function Theorem in Section II.3 of these notes implies that, if we restrict to a small enough region U containing (a, b), then the set of points in U satisfying $F(x, y) = 0$ is equal to the graph of some function $y = h(x)$ if the first partial derivative of F at (a, b) is nonzero. Similarly, if the second partial derivative of F at (a, b) is nonzero, then there is a small region U containing (a, b) such that the set of points in U satisfying $F(x, y) = 0$ is equal to the graph of some function $x = k(y)$. If we combine the conclusions in the preceding sentences, we may conclude that the set of points satisfying $F(x, y) = 0$ can be split into pieces such that each has a smooth parametrization. The ordinary unit circle defined by $x^2 + y^2 = 1$ is an example of a curve that is near some points as the graph of a function of x and near other points as the graph of a function of y, but cannot be expressed globally as the graph of a function of either x or y (for example, if it were globally the graph of a function of x then every vertical line defined by an equation of the form $x = c$ would meet the curve in at most one point, and clearly there are many values of c for which the curve meets the vertical line in two points).

To indicate the importance of describing curves as sets of points (x, y) such that $F(x, y) = 0$, we need only recall that one way of characterizing lines and conics in the plane is that lines in the planes are the curves whose coordinates (x, y) satisfy a nontrivial first degree polynomial equation $p(x, y) = 0$, and conics are the curves which satisfy a nontrivial polynomial equation $p(x, y) = 0$ such that the polynomial p has degree 2. More generally, one can define a plane curve to be *algebraic* if its coordinates satisfy a nontrivial polynomial equation $p(x, y) = 0$, and similarly a curve is *transcendental* if there is no nonzero polynomial p whose coordinates satisfy the equation $p(x, y) = 0$. A discussion of algebraic and transcendental curves appears in the following online documents:

`http://math.ucr.edu/~res/math153/transcurves.pdf`

`http://math.ucr.edu/~res/math153/transcurves2.pdf`

`http://math.ucr.edu/~res/math153/transcurves3.pdf`

Addendum: Note on smoothness classes

We shall construct examples of real valued functions $f(x)$ which have continuous derivatives of order $\leq n$ but no derivative of order $n+1$. If $n = 0$ then the function $f(x) = |x|$ (absolute value) is such an example because f is continuous but $f'(0)$ cannot be defined. Examples for higher values of n are obtained recursively from this example.

Start with $f(x) = x\,|x|$. Since $f(x) = x^2$ if $x > 0$ and $f(x) = -x^2$ if $x < 0$, it follows quickly that $f'(x) = 2\,|x|$ if $x \neq 0$. Furthermore, the definition of derivative implies that $f'(0) = 0$, so f has a continuous derivative everywhere. It also follows that if $x \neq 0$ then $f''(x) = \mathrm{sign}(x)$, and since $f'(x) = 2\,|x|$ it follows that $f''(0)$ cannot be defined.

Now let $f_n(x) = x^n|x|$ for $n \geq 1$. We claim that f_n has continuous derivatives of order n but $f^{(n+1)}(0)$ cannot be defined. The key step is to prove that $m \geq 2$ implies

$$f'_m(x) \quad = \quad (m+1)\, f_{m-1}(x)\ , \qquad \text{for all} \quad x\ .$$

This follows from the identity $f_m(x) = x^{m-1}\, f_1(x)$ and the usual rule for differentiating products. Similarly, if we differentiate repeatedly we obtain the formula

$$\frac{d}{dx}^{\,m} f_m(x) \quad = \quad (m+1)!\,|x|\ .$$

In particular, this means that f_m has continuous derivatives of all orders $\leq m$, but $f^{(m+1)}(0)$ cannot be defined because $|x|$ has no derivative at $x = 0$.■

I.3 : Arc length and reparametrization

(Lipschutz, Chapter 3)

Given a parametrized smooth regular curve $\mathbf{x}$ defined on a closed interval $[a, b]$, as in calculus we define the *arc length* of $\mathbf{x}$ from $t = a$ to $t = b$ to be the integral

$$L \quad = \quad \int_a^b |\mathbf{x}'(t)|\, dt \ .$$

The motivation for this definition is usually discussed in calculus courses, and it is reviewed below in the subsection on arc length for curves that are not necessarily smooth. More generally, if $a \leq t \leq b$ then the length of the curve from parameter value a to parameter value t is given by

$$s(t) \quad = \quad \int_a^t |\mathbf{x}'(u)|\, du \ .$$

By the Fundamental Theorem of Calculus, the partial arc length function s is differentiable on $[a, b]$ and its derivative is equal to $|\mathbf{x}'(t)|$. If we have a regular smooth curve, this function is continuous and everywhere positive (hence $s(t)$ is a strictly increasing function of t), and the image of this function is equal to the closed interval $[0, L]$.

COMPUTATIONAL ISSUES. Although the arc length formula is fairly simple to state, it can be extremely difficult to evaluate the integrals which it yields, even for familiar curves with relatively simple parametrizations. For example, if one applies the formula to an arc on an ellipse, then results of J. Liouville from the nineteenth century show that one cannot express the resulting integral in terms of the standard functions considered in first year calculus. Here are some further references:

`http://en.wikipedia.org/wiki/Elliptic_integral`

`http://math.ucr.edu/~res/math10B/nonelementary_integrals.pdf`

The result about arc length for ellipses can be derived from the material in the first reference (combined with material in J. F. Ritt, *Integration in finite terms*, Columbia Univ. Press, 1948, pp. 35–37). Another example of an "impossible" arc length integral is given in Section I.4 below (see the subheading *Computational techniques*).

Reparametrizations of curves

Given a parametrized curve $\mathbf{x}$ defined on an interval $[a, b]$, it is easy to find other parametrizations by simple changes of variables. For example, the curve $\mathbf{y}(t) = \mathbf{x}(t + a)$ resembles the original curve in many respects: For example, both have the same tangent vectors and images, and the only real difference is that $\mathbf{y}$ is defined on $[0, b - a]$ rather than $[a, b]$. Less trivial changes of variable can be extremely helpful in analyzing the image of a curve. For example, the parametrized curve $\mathbf{x}(t) = (e^t - e^{-t},\ e^t + e^{-t})$ has the same image as the the upper branch of the hyperbola $y^2 - x^2 = 4$ (*i.e.*, the graph of $y = \sqrt{4+x^2}$); as a graph, this curve can also be parametrized using the graph parametrization $\mathbf{y}(u) = (u,\ \sqrt{4+u^2})$. These parametrizations are related by the change of variables $u = 2\sinh t$; in other words, we have

$$\mathbf{x}(t) = \mathbf{y}\big(\,2\sinh t\,\big)\ .$$

Note that u varies from $-\infty$ to $+\infty$ as t goes from $-\infty$ to $+\infty$, and $u'(t) = \cosh t\ >\ 0$ for all t.

More generally, it is useful to consider reparametrizations of curves corresponding to functions $u(t)$ such that $u'(t)$ is never zero. Of course the sign of u' determines whether u is strictly increasing or decreasing, and it is useful to allow both possibilities. Suppose that we are given a differentiable function u defined on $[a,b]$ such that u' is never zero on $[a,b]$. Then the image ot u is some other closed interval, say $[c,d]$; if u is increasing then $u(a) = c$ and $u(b) = d$, while if u is decreasing then $u(a) = d$ and $u(b) = c$. It follows that u has an inverse function t defined on $[c,d]$ and taking values in $[a,b]$. Furthermore, the derivatives dt/du and du/dt are reciprocals of each other by the standard formula for the derivative of an inverse function.

It is important to understand how reparametrization changes geometrical properties of a curve such as tangent lines and arc lengths. The most basic thing to consider is the effect on tangent vectors.

PROPOSITION. *Let* $\mathbf{x}$ *be a regular smooth curve defined on the closed interval* $[c,d]$*, let* $u : [a,b] \to [c,d]$ *be a function with a continuous derivative that is nowhere zero, and let* $\mathbf{y}(t) = \mathbf{x}\big(\,u(t)\,\big)$*. Then*

$$\mathbf{y}'(t) \quad = \quad u'(t)\cdot\mathbf{x}'\big(\,u(t)\,\big)\ .$$

This is an immediate consequence of the Chain Rule.■

COROLLARY. *For each* $t \in [a,b]$ *the tangent line to* $\mathbf{y}$ *at parameter value* t *is the same as the tangent line to* $\mathbf{x}$ *at* $u(t)$*. Furthermore, the standard parametrizations are related by a linear change of coordinates.*

Proof. By definition, the tangent line to $\mathbf{x}$ at $u(t)$ is the line joining $\mathbf{x}\big(\,u(t)\,\big)$ and $\mathbf{x}\big(\,u(t)\,\big) + \mathbf{x}'\big(\,u(t)\,\big)$. Similarly, the tangent line to $\mathbf{y}$ at t is the line joining $\mathbf{y}(t) = \mathbf{x}\big(\,u(t)\,\big)$ and

$$\mathbf{y}(t) + \mathbf{y}'(t) = \mathbf{x}\big(\,u(t)\,\big) + u'(t)\,\mathbf{x}'\big(\,u(t)\,\big)\ .$$

Since the line joining the distinct points (or vectors) $\mathbf{a}$ and $\mathbf{a}+\mathbf{b}$ is the same as the line joining $\mathbf{a}$ and $\mathbf{a} + c\,\mathbf{b}$ if $c \neq 0$, it follows that the two tangent lines are the same (take $\mathbf{a} = \mathbf{y}(t)$, $\mathbf{b} = \mathbf{x}'(u)$ and $c = u'(t)$).

In fact, we have obtained standard linear parametrizations of this line given by $\mathbf{f}(z) = \mathbf{a} + z\,\mathbf{b}$ and $\mathbf{g}(w) = \mathbf{a} + cw\,\mathbf{b}$. It follows that $\mathbf{g}(w) = \mathbf{f}(cw)$.■

Arc length is another property of a curve that does not change under reparmetrization.

PROPOSITION. *Let* $\mathbf{x}$ *be a regular smooth curve defined on the closed interval* $[c,d]$*, let* $u : [a,b] \to [c,d]$ *be a function with a continuous derivative that is nowhere zero, and let* $\mathbf{y}(t) = \mathbf{x}\big(\,u(t)\,\big)$*. Then*

$$\int_c^d |\mathbf{x}'(u)|\,du = \int_a^b |\mathbf{y}'(t)|\,dt$$

Proof. The standard change of variables formula for integrals implies that

$$\int_c^d |\mathbf{x}'(u)|\, du = \int_a^b |\mathbf{x}'\big(\, u(t)\, \big)\, |u'(t)|\, dt\ .$$

Some comments about this formula and the absolute value sign may be helpful. If u is increasing then the sign is positive and we have $u(a) = c$ and $u(b) = d$, so $|u'(t)| = u'(t)$; on the other hand if u is decreasing, then the Fundamental Theorem of Calculus suggests that the integral on the left hand side should be equal to

$$\int_b^a \left|\mathbf{x}'\big(\, u(t)\, \big)\right| \cdot u'(t)\, dt \quad = \quad -\int_a^b \left|\mathbf{x}'\big(\, u(t)\, \big)\right| \cdot u'(t)\, dt \quad = \quad \int_a^b \left|\mathbf{x}'\big(\, u(t)\, \big)\right| \cdot [-u'(t)]\ dt$$

so that the formula above holds because $u' < 0$ implies $|u'| = -u'$. In any case, the properties of vector length imply that the integrand on the right hand side of the change of variables equation is $|u'(t) \cdot \mathbf{x}'(u)|$, which by the previous proposition is equal to $|\mathbf{y}'(t)|$.■

If $\mathbf{v}$ is a regular smooth curve defined on $[a, b]$, then the arc length function

$$s(t) = \int_a^t |\mathbf{v}'(u)|\, du$$

often provides an extremely useful reparametrization because of the following result:

PROPOSITION. *Let $\mathbf{v}$ be as above, and let $\mathbf{x}$ be the reparametrization defined by $\mathbf{x}(s) = \mathbf{v}\big(\mu(s)\big)$, where μ is the inverse function to the arc length function $\lambda : [a, b] \to [0, L]$. Then $|\mathbf{x}'(s)| = 1$ for all s.*

Proof. By the Fundamental Theorem of Calculus we know that $\lambda'(t) = |\mathbf{v}'(t)|$. Therefore by the Chain Rule we know that

$$\mathbf{x}'(s) \quad = \quad \mu'(s)\, \mathbf{v}'\big(\, \mu(s)\, \big)$$

and by the differentiation formula for inverse functions we know that

$$\mu'(s) \quad = \quad \frac{1}{\lambda'(\, \mu(s)\,)} \quad = \quad T'(s) \quad = \quad \frac{1}{|\mathbf{v}'(\, T(s)\,)|}$$

and if we substitute this into the expression given by the chain rule we see that

$$|\mathbf{x}'(s)| \quad = \quad |T'(s)|\, |\mathbf{v}'\big(\, T(s)\, \big)| \quad = \quad \frac{1}{|\mathbf{v}'\big(\, T(s)\, \big)|} \cdot |\mathbf{v}'\big(\, T(s)\, \big)| \quad = \quad 1\ .■$$

Arc length for more general curves

The geometric motivation for the definition of arc length is described in Exercises 8–0 on pages 10–11 of DO CARMO; specifically, given a parametrized curve $\mathbf{x}$ defined on $[a, b]$ one picks a finite set of points t_i such that

$$a \quad = \quad t_0 \ < \ t_1 \ < \ \cdots \ < \ t_m = b$$

and views the length of the inscribed broken line joining t_0 to t_1, t_1 to t_2 *etc.* as an approximation to the length of the curve. In favorable circumstances if one refines the finite set of points by taking more and more of them and making them closer and closer together, the lengths of these broken line curves will have a limiting value which is the arc length. Exercise 9(b) on page 11 of DO CARMO gives one example of a curve for which no arc length can be defined. During the time since do Carmo's book and the Schaum's Outline Series book were published, a special class of such curves known as *fractal curves* has received a great deal of attention. The parametric equations defining such curves all have the form $\mathbf{x}(t) = \lim_{n\to\infty} \mathbf{x}_n(t)$, where each $\mathbf{x}_n$ is a piecewise smooth regular curve and for each n one obtains $\mathbf{x}_n$ from $\mathbf{x}_{n-1}$ by making some small but systematic changes. Some online references with more information on such curves are given below.

`http://mathworld.wolfram.com/Fractal.html`

`http://ecademy.agnesscott.edu/~lriddle/ifs/ksnow/lsnow/htm`

`http://en2.wikipedia.org/wiki/Koch_snowflake`

`http://en.wikipedia.org/wiki/Fractal_geometry`

`http://www.youtube.com/watch?v=a9xvz_Palg`

The preceding discussion illustrates that parametrized curves include a wide range of objects which do not have piecewise smooth reparametrizations. However, thus far the images of the parametrizations "look like" the standard examples in some vague sense; namely, they are topologically equivalent to intervals in the real line or ordinary circles in the plane. It is also possible to find even more bizarre examples of parametrized curves. In particular, one can construct parametrized curves whose image is the entire coordinate plane or 3-space. Here is an online reference for such space-filling curves:

`http://en.wikipedia.org/wiki/Space-filling_curve`

`http://en.wikipedia.org/wiki/Hilbert_curve`

A more formal account appears in Section 44 of the following graduate level textbook:

J. R. Munkres. Topology. (Second Edition), *Prentice-Hall, Saddle River NJ*, 2000. ISBN: 0–13–181629–2.

I.4 : Curvature and torsion

(Lipschutz, Chapter 4)

Many calculus courses include a brief discussion of curvature, but the approaches vary and it will be better to make a fresh start.

Definition. Let $\mathbf{x}$ be a regular smooth curve, and assume it is parametrized by arc length plus a constant (*i.e.*, $|\mathbf{x}'(s)| = 1$ for all s). The *curvature* of $\mathbf{x}$ at parameter value s is equal to $\kappa(s) = |\mathbf{x}''(s)|$.

The most immediate question about this definition is why it has anything to do with our intuitive idea of curvature. The best way to answer this is to look at some examples.

Suppose that we are given a parametrized line with an equation of the form $\mathbf{x}(t) = \mathbf{a} + t\,\mathbf{b}$ where $|\mathbf{b}| = 1$. It then follows that $\mathbf{x}$ is parametrized by arc length by means of t, and clearly we have $\mathbf{x}''(t) = \mathbf{0}$. This means that the curvature of the line is zero at all points, which is what we expect.

Consider now an example that is genuinely curved; namely, the circle of radius r about the origin. The arc length parametrization for this curve has the form

$$\mathbf{x}(s) \quad = \quad \Big(r\cos(s/r),\, r\sin(s/r)\Big)$$

and one can check directly that its first two derivatives are given as follows:

$$\mathbf{x}'(s) \quad = \quad \Big(-\sin(s/r),\, \cos(s/r)\Big)$$

$$\mathbf{x}''(s) \quad = \quad \left(-\frac{\cos(s/r)}{r},\, -\frac{\sin(s/r)}{r}\right)$$

It follows that *the curvature of the circle at all points is given by the reciprocal of the radius.*■

The following simple property of the "acceleration" function $\mathbf{x}''(s)$ turns out to be quite important for our purposes:

PROPOSITION. *The vectors $\mathbf{x}''(s)$ and $\mathbf{x}'(s)$ are perpendicular.*

Proof. We know that $|\mathbf{x}'(s)|$ is always equal to 1, and thus the same is true of its square, which is just the dot product of $\mathbf{x}'(s)$ with itself. The product rule for differentiating dot products of two functions then implies that

$$0 \quad = \quad \frac{d}{ds}\Big(\mathbf{x}'(s)\cdot\mathbf{x}'(s)\Big) \quad = \quad 2\left(\mathbf{x}'(s)\cdot\mathbf{x}''(s)\right)$$

and therefore the two vectors are indeed perpendicular.■

Geometric interpretation of curvature

We begin with a very simple observation.

PROPOSITION. *If $\mathbf{x}(s)$ is a smooth curve (parametrized by arc length) whose curvature $\kappa(s)$ is zero for all s, then $\mathbf{x}(s)$ is a straight line curve of the form $\mathbf{x}(s) = \mathbf{x}(0) + s\,\mathbf{x}'(0)$.*

Proof. Since $\kappa(s)$ is the length of $\mathbf{x}''(s)$, if the curvature is always zero then the same is true for $\mathbf{x}''(s)$. But this means that $\mathbf{x}'(s)$ is constant and hence equal to $\mathbf{x}'(0)$ for all s, and the latter in turn implies that $\mathbf{x}(s) = \mathbf{x}(0) + s\,\mathbf{x}'(0)$.■

Given a smooth curve, the tangent line to the curve at a point t may be viewed as a first order linear approximation to the curve. The notion of curvature is related to a corresponding second order approximation to the curve at parameter value t by a line or circle. We begin by making this notion precise:

Defintion. Let n be a positive integer. Given two curves $\mathbf{a}(t)$ and $\mathbf{b}(t)$ defined on an interval J containing t_0 such that $\mathbf{a}(t_0) = \mathbf{b}(t_0)$, we say that $\mathbf{a}$ *and* $\mathbf{b}$ *are strong n^{th} order approximations to each other* if there is an $\varepsilon > 0$ such that $|h| < \varepsilon$ and $t_0 + h \in J$ imply

$$|\mathbf{b}(t_0+h) \;-\; \mathbf{a}(t_0+h)| \quad \leq \quad C\,|h|^{n+1}$$

for some constant $C > 0$. The analytic condition on the order of approximation is often formulated geometrically as the order of contact that two curves have with each other at a given point; as the order of contact increases, so does the speed at which the curves approach each other. The most basic visual examples here are the x-axis and the graphs of the curves x^n near the origin. Further information relating geometric ideas of high order contact and Taylor polynomial approximations is presented on pages 87–91 of the Schaum's Outline Series book on differential geometry (bibliographic information is given at the beginning of these notes).

LEMMA. *Suppose that the curves $\mathbf{a}(t)$ and $\mathbf{b}(t)$ are defined on an interval J containing t_0 such that $\mathbf{a}(t_0) = \mathbf{b}(t_0)$, and assume also that $\mathbf{a}$ and $\mathbf{b}$ are strong n^{th} order approximations to each other at t_0. Then for each regular smooth reparametrization $t(u)$ with $t_0 = t(u_0)$ the curves $\mathbf{a}\circ t$ and $\mathbf{b}\circ t$ are strong n^{th} order approximations to each other at u_0.*

Proof. Let J_0 be the domain of the function $t(u)$, and let K_0 be a closed bounded subinterval containing u_0 such that the latter is an endpoint of K_0 if and only if it is an endpoint of J_0. Denote the maximum value of $|t'(u)|$ on this interval by M. Then by hypothesis and the Mean Value Theorem we have

$$|\mathbf{b}\big(t(u_0+h)\big) \;-\; \mathbf{a}\big(t(u_0+h)\big)| \quad\leq\quad C\,|t(u_0+h) - t(u_0)|^{n+1} \quad\leq\quad C\,M^{n+1}\cdot|h|^{n+1}$$

which proves the assertion of the lemma.■

In the terminology of n^{th} order approximations, if we are given a regular smooth curve $\mathbf{x}$ then a strong first order approximation to it is given by the tangent line with the standard linear parametrization

$$\mathbf{L}(t_0+h) = \mathbf{x}(t_0) + h\,\mathbf{x}'(t)\ .$$

Furthermore, this line is the unique strong first order linear approximation to $\mathbf{x}$.

Here is the main result on curvature and strong second order approximations.

THEOREM. *Let $\mathbf{x}$ be a regular smooth curve defined on an interval J containing 0 such that $\mathbf{x}'$ has a continuous* **second** *derivative and $|\mathbf{x}'| = 1$ (hence $\mathbf{x}$ is parametrized by arc length plus a constant).*

(i) *If the curvature of $\mathbf{x}$ at 0 is zero, then the tangent line is a strong second order approximation to $\mathbf{x}$.*

(ii) *Suppose that the curvature of $\mathbf{x}$ at 0 is nonzero, let $\mathbf{N}$ be the unit vector pointing in the same direction as $\mathbf{x}''(0)$ (the latter is nonzero by the definition and nonvanishing of the curvature at parameter value 0). If Γ is the circle through $\mathbf{x}(0)$ such that [1] its center is $\mathbf{x}(0) + (\kappa(0))^{-1}\mathbf{N}$, [2] it lies in the plane containing this center and the tangent line to the curve at parameter value zero, then Γ is a strong second order approximation to $\mathbf{x}$.*

For the sake of completeness, we shall describe the unique plane containing a given line and an external point explicitly as follows. If $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are noncollinear points in $\mathbb{R}^3$ then the plane containing them consists of all $\mathbf{x}$ such that $\mathbf{x} - \mathbf{a}$ is perpendicular to

$$\big(\mathbf{b} - \mathbf{a}\big) \;\times\; \big(\mathbf{c} - \mathbf{a}\big)$$

which translates to the triple product equation

$$[(\mathbf{x} - \mathbf{a}),\, (\mathbf{b} - \mathbf{a}),\, (\mathbf{c} - \mathbf{a})] \quad=\quad 0\ .$$

Suppose now that $\mathbf{b}_1$ and $\mathbf{c}_1$ are points on the line containing $\mathbf{b}$ and $\mathbf{c}$. Then we may write

$$\mathbf{b}_1 \;=\; u\,\mathbf{b} \;+\; (1-u)\,\mathbf{c}\;, \quad \mathbf{c}_1 \;=\; v\,\mathbf{b} \;+\; (1-v)\,\mathbf{c}$$

for suitable real numbers u and v. The equations above immediately imply the following identities:

$$(\mathbf{b}_1 \;-\; \mathbf{a}) \;=\; u\,(\mathbf{b} \;-\; \mathbf{a}) \;+\; (1-u)\,(\mathbf{c} \;-\; \mathbf{a}$$

$$(\mathbf{c}_1 \;-\; \mathbf{a}) \;=\; v\,(\mathbf{b} \;-\; \mathbf{a}) \;+\; (1-v)\,(\mathbf{c} \;-\; \mathbf{a})\;.$$

These formulas and the basic properties of determinants imply

$$[(\mathbf{x}-\mathbf{a}).\,(\mathbf{b}_1-\mathbf{a}),\,(\mathbf{c}_1-\mathbf{a})] \;=$$

$$[(\mathbf{x}-\mathbf{a}).\,u(\mathbf{b}_1-\mathbf{a}),\,v(\mathbf{c}_1-\mathbf{a})] \;+\; \big[(\mathbf{x}-\mathbf{a}).\,(1-u)\big(\mathbf{b}_1-\mathbf{a}\big),\,(1-v)\big(\mathbf{c}_1-\mathbf{a}\big)\big] \;=$$

$$uv\,[(\mathbf{x}-\mathbf{a}),\,(\mathbf{b}-\mathbf{a}),\,(\mathbf{c}-\mathbf{a})] \;+\; (1-u)\,(1-v)\,[(\mathbf{x}-\mathbf{a}),\,(\mathbf{c}-\mathbf{a}),\,(\mathbf{b}-\mathbf{a})] \;=$$

$$uv\,0 \;-\; (1-u)\,(1-v)\,0 \;=\; 0$$

and hence the equation

$$[(\mathbf{x}-\mathbf{a}),\,(\mathbf{b}-\mathbf{a}),\,(\mathbf{c}-\mathbf{a})] \;=\; 0$$

implies the corresponding equation if $\mathbf{b}$ and $\mathbf{c}$ are replaced by two arbitrary points on the line containing $\mathbf{b}$ and $\mathbf{c}$.■

Proof of Proposition. Consider first the case where $\kappa(0) = 0$. Then the tangent line to the curve has equation $\mathbf{L}(s) = s\,\mathbf{x}'(0)$ and the second order Taylor expansion for $\mathbf{x}$ has the form $\mathbf{x}(s) = s\,\mathbf{x}'(0) + \frac{1}{2}s^2\,\mathbf{x}''(0) + s^3\,\theta(s)$ where $\theta(s)$ is bounded for s sufficiently close to zero. The assumption $\kappa(0) = 0$ implies that $\mathbf{x}''(0) = 0$ and therefore we have $\mathbf{x}(s) - \mathbf{L}(s) = s^3\,\theta(s)$ where $\theta(s)$ is bounded for s sufficiently close to zero. Therefore the tangent line is a strong second order approximation to the curve if the curvature is equal to zero.

Suppose now that $\kappa(0) \neq 0$, and let $\mathbf{N}$ be the unit vector pointing in the same direction as $\mathbf{x}''(0)$. Define $\mathbf{z}$ by the formula

$$\mathbf{z} \;=\; \mathbf{x}(0) \;+\; \frac{1}{\kappa(0)}\,\mathbf{N}$$

and consider the circle in the plane of $\mathbf{z}$ and the tangent line to $\mathbf{x}$ at parameter value $s = 0$ such that the center is $\mathbf{z}$ and the radius is $1/\kappa(0)$. If we set r equal to $1/\kappa(0)$ and $\mathbf{T} = \mathbf{x}'(0)$, then a parametrization of this circle in terms of arc length is given by

$$\Gamma(s) \;=\; \mathbf{z} \;-\; r\cos(s/r)\,\mathbf{N} \;+\; r\sin(s/r)\,\mathbf{T}\;.$$

Using the standard power series expansions for the sine and cosine function and the identity $\mathbf{z} = \mathbf{x}(0) - r\,\mathbf{N}$, we may rewrite this in the form

$$\Gamma(s) \;=\; \mathbf{x}(0) \;+\; \frac{s^2}{2r}\mathbf{N} \;+\; s^3\alpha(s)\mathbf{N} + s\mathbf{T} + s^3\beta(s)\mathbf{T}$$

where $\alpha(s)$ and $\beta(s)$ are continuous functions and hence are bounded for s close to zero. On the other hand, using the Taylor expansion of $\mathbf{x}(s)$ near $s = 0$ we may write $\mathbf{x}(s)$ in the form

$$\mathbf{x}(0) \;+\; s\,\mathbf{x}'(0) \;+\; \frac{s^2}{2}\mathbf{x}''(0) \;+\; s^3\,\mathbf{W}(s)$$

where $\mathbf{W}(s)$ is bounded for s close to zero. But $\mathbf{x}'(0) = \mathbf{T}$ and

$$\mathbf{x}''(0) \quad = \quad \kappa(0)\,\mathbf{N} \quad = \quad \frac{1}{r}\mathbf{N}$$

so that $\Gamma(s) - \mathbf{x}(s)$ has the form $s^3\mathbf{W}_1(s)$ where $\mathbf{W}_1(s)$ is a bounded function of s. Therefore the circle defined by Γ is a strong second order approximation to the original curve at the parameter value $s = 0$.■

Notation. If the curvature of $\mathbf{x}$ is nonzero near parameter value s as in the proposition, then the center of the strong second order circle approximation

$$\mathbf{z}(s) \quad = \quad \mathbf{x}(s) \;+\; \frac{1}{(\kappa(s))^2}\,\mathbf{x}''(s)$$

is called the *center of curvature* of $\mathbf{x}$ at parameter value s. The circle itsef is called the *osculating circle* to the curve at parameter value s (in Latin, *osculare* = to kiss).

Complementary result. A more detailed analysis of the situation shows that if $\kappa(0) \neq 0$ then the circle given above is the unique circle that is a second order approximation to the original curve at the given point.■

Computational techniques

Although the description of curvature in terms of arc length parametrizations is important for theoretical purposes, it is usually not particularly helpful if one wants to compute the curvature of a given curve at a given point. One major reason for this is that the arc length function $s(t)$ can only be written down explicitly in a very restricted class of cases. In particular, if we consider the graph of the cubic polynomial $y = x^3$ with parametrization $(t,\, t^3)$ on some interval $[0, a]$ then the arc length parameter is given by the formula

$$s(t) \quad = \quad \int_0^t \sqrt{1 + 9\,u^4}\,du$$

and results of P. Chebyshev from the nineteenth century show that there is no "nice" formula for this function in terms of the usual functions one studies in first year calculus. Therefore it is important to have formulas for curvature in terms of arbitrary parametrizations of a regular smooth curve.

Remarks.

1. The statement about the antiderivative of $\sqrt{1 + 9\,x^4}$ is stronger than simply saying that no one has has been able to find a nice formula for the antiderivative. It as just as impossible to find one as it is to find two positive whole numbers a and b such that $\sqrt{2} = a/b$ or to find two even positive integers whose sum is an odd integer.

2. A detailed statement of Chebyshev's result can be found on the web link

`http://mathworld.wolfram.com/IndefiniteIntegral.html`

and further references are also given there.

The following formula appears in many calculus texts:

FIRST CURVATURE FORMULA. *Let* $\mathbf{x}$ *be a smooth regular curve, let* s *be the arc length function, let* $k(t) = \kappa(\,s(t)\,)$*, and let* $\mathbf{T}(t)$ *be the unit tangent vector function obtained by multiplying* $\mathbf{x}'(t)$ *by the reciprocal of its length. Then we have*

$$k(t) \quad = \quad \frac{|\mathbf{T}\,'(t)|}{|\mathbf{x}'(t)|} \ .$$

Derivation. We have seen that $\mathbf{T}\,(s)$ is equal to $\mathbf{x}'(s)$, and therefore by the chain rule we have

$$\mathbf{T}\,'(t) \quad = \quad s'(t)\,\mathbf{T}\,'(\,s(t)\,) \quad = \quad |\mathbf{x}'(t)|\,\mathbf{x}''(s) \ .$$

Taking lengths of the vectors on both sides of this equation we see that

$$|\mathbf{T}\,'(t)| \quad = \quad |\mathbf{x}'(t)| \cdot |\mathbf{x}''(s)| \quad = \quad |\mathbf{x}'(t)|\,k(t)$$

which is equivalent to the formula for $k(t)$ displayed above.∎

Here is another formula for curvature that is often found in calculus textbooks.

SECOND CURVATURE FORMULA. *Let* $\mathbf{x}$ *be a smooth regular curve, let* s *be the arc length function, let* $\mathbf{T}(t)$ *be the unit length tangent vector function, and let* $k(t) = \kappa(\,s(t)\,)$*. Then we have*

$$k(t) \quad = \quad \frac{|\mathbf{x}'(t) \ \times \ \mathbf{x}''(t)|}{|\mathbf{x}'(t)|^3} \ .$$

Derivation. As in the derivation of the First Curvature Formula we have $\mathbf{x}' = s'\mathbf{T}$. Therefore the Leibniz product rule for differentiating the product of a scalar function and a vector function yields

$$\mathbf{x}'' \quad = \quad s''\mathbf{T} \ + \ s'\mathbf{T}\,' \ .$$

Since $\mathbf{T} \times \mathbf{T} = \mathbf{0}$ the latter implies

$$\mathbf{x}' \times \mathbf{x}'' \quad = \quad (s')^2 \, \left(\mathbf{T} \times \mathbf{T}\,'\right) \ .$$

Since $|\mathbf{T}| = 1$ it follows that $\mathbf{T} \cdot \mathbf{T}\,' = 0$; *i.e.*, the vectors $\mathbf{T}$ and $\mathbf{T}\,'$ are orthogonal. This in turn implies that $|\mathbf{T} \times \mathbf{T}\,'|$ is equal to $|\mathbf{T}| \cdot |\mathbf{T}\,'|$ so that

$$|\mathbf{x}' \times \mathbf{x}''| \quad = \quad |s'|^2\,|\mathbf{T} \times \mathbf{T}\,'| \quad = \quad |s'|^2\,|\mathbf{T}| \cdot |\mathbf{T}\,'| \quad = \quad (s')^2\,|\mathbf{T}\,'| \quad = \quad |\mathbf{x}'|^2\,|\mathbf{T}\,'|$$

(at the next to last step we again use the identity $|\mathbf{T}| = 1$). It follows that

$$|\mathbf{T}\,'| \quad = \quad \frac{|\mathbf{x}'(t) \ \times \ \mathbf{x}''(t)|}{|\mathbf{x}'(t)|^2}$$

and the Second Curvature Formula follows by substitution of this expression into the First Curvature Formula.∎

Computations of curvature for some familiar examples of curves are given in the file `curvexamples.pdf`.

Thus far we have discussed lines and circles that are good approximations to a curve. Given a curve in 3-dimensional space one can also ask whether there is some plane that comes as close as possible to containing the given curve. Of course, for curves that lie entirely in a single plane, the definition should yield this plane.

Given a continuous curve $\mathbf{x}(t)$, and a plane Π, one way of making this notion precise is to consider the function $\Delta(t)$ giving the distance from $\mathbf{x}(t)$ to Π. If the point $\mathbf{x}(t_0)$ lies on Π, then $\Delta(t_0) = 0$ and one test of how close the curve comes to lying in the plane is to determine the extent to which the zero function is an n^{th} order approximation to $\Delta(t)$ for various choices of n. In fact, if $\kappa(t_0) \neq 0$ then there is a unique plane such that the zero function is a second order approximation to $\Delta(t)$, and this plane is called the *osculating plane* to $\mathbf{x}$ at parameter value $t = t_0$. Formally, we proceed as follows:

Definition. Let $\mathbf{x}(s)$ be a regular smooth curve parametrized by arc length (so that $|\mathbf{x}'| = 1$), and assume that $\kappa(s_0) \neq 0$. Let $\mathbf{a} = \mathbf{x}(0)$, let $\mathbf{T} = \mathbf{x}'(s_0)$, and let $\mathbf{N}$ be the unit vector pointing in the same direction as $\mathbf{x}''(s_0)$. The *osculating plane* to the curve at parameter value s_0 is the unique plane containing the three noncollinear vectors $\mathbf{a}$, $\mathbf{a} + \mathbf{T}$, and $\mathbf{a} + \mathbf{N}$.

It follows that the equation defining the osculating plane may be written in the form

$$[(\mathbf{y} - \mathbf{a}),\, \mathbf{T},\, \mathbf{N}] \quad = \quad 0 \ .$$

We can now state the result on the order of contact between curves and their osculating planes.

PROPOSITION. *Let $\mathbf{x}$ be a regular smooth curve parametrized by arc length (hence $|\mathbf{x}'| = 1$), assume that $\mathbf{x}$ has a continuous third derivative, and assume also that $\kappa(s_0) \neq 0$. Let Π be the osculating plane of $\mathbf{x}$ at parameter value s_0, and let $\Delta(s)$ denote the distance between $\mathbf{x}(s)$ and Π. Then the osculating plane is the unique plane through $\mathbf{x}(s_0)$ such that the zero function is a second order approximation to the distance function $\Delta(s)$ at s_0.*

Proof. Let $\mathbf{a} = \mathbf{x}(s_0)$, let $\mathbf{T} = \mathbf{x}'(s_0)$, let $\mathbf{N}$ be the unit vector pointing in the same direction as $\mathbf{x}''(s_0)$, and let $\mathbf{B}$ be the cross product $\mathbf{T} \times \mathbf{N}$. Then the oscularing plane is the unique plane containing $\mathbf{a}$, $\mathbf{a} + \mathbf{T}$, and $\mathbf{a} + \mathbf{N}$, and the distance between a point $\mathbf{y}$ and the osculating plane is the absolute value of the function $\widetilde{D}\,(\mathbf{y}) = (\mathbf{y} - \mathbf{a}) \cdot \mathbf{B}$. The second order Taylor approximation to $\mathbf{x}(s)$ with respect to s_0 is then given by the formula

$$\mathbf{x}(s) = \mathbf{a} \ + \ (s - s_0) \cdot \mathbf{T} \ + \ \frac{(s - s_0)^2\, \kappa(s_0)}{2} \cdot \mathbf{N} \ + \ (s - s_0)^3\, \mathbf{W}(s)$$

where $\mathbf{W}(s)$ is bounded for s sufficiently close to s_0. Therefore since $\mathbf{B}$ is perpendicular to $\mathbf{T}$ and $\mathbf{N}$ we have

$$\widetilde{D}\,\big(\mathbf{x}(s))\big) \quad = \quad (s - s_0)^3\, \mathbf{W}(s) \cdot \mathbf{B}$$

where $\mathbf{W}(s) \cdot \mathbf{B}$ is bounded for s sufficiently close to s_0. Therefore the given curve has order of contact at least two with respect to its osculating plane.

Suppose now that we are given some other plane through $\mathbf{a}$; then one has a normal vector $\mathbf{V}$ to the plane of the form $\mathbf{B} + p\,\mathbf{T} + q\,\mathbf{N}$ where p and q are not both zero. The distance between $\mathbf{x}(s)$ and plane through $\mathbf{a}$ with normal vector $\mathbf{V}$ will then be the absolute value of a nonzero multiple of the function

$$\Big(\, (\mathbf{x}(s) \ - \ \mathbf{a}) \cdot \mathbf{V} \Big)$$

which is equal to

$$g(s-s_0) \quad = \quad (s-s_0)\big(\mathbf{T}\cdot\mathbf{V}\big) \ + \ \frac{(s-s_0)^2\,\kappa(s_0)}{2}\,\big(\mathbf{N}\cdot\mathbf{V}\big) \ + \ (s-s_0)^3\,\big(\mathbf{W}(s)\cdot\mathbf{V}\big)\ .$$

We then have

$$\frac{g(s-s_0)}{(s-s_0)^3} \quad = \quad \frac{p}{(s-s_0)^2} \ + \ \frac{q}{(s-s_0)} \ + \ \big(\mathbf{W}(s)\cdot\mathbf{V}\big)$$

where the third term on the right is bounded. But since at least one of p and q is nonzero, it follows that the entire sum is not a bounded function of s if s is close to s_0. Therefore the curve cannot have order of contact at least two with any other plane through **a**.■

Torsion

Curvature may be viewed as reflecting the rate at which a curve moves off its tangent line. The notion of torsion will reflect the rate at which a curve moves off its osculating plane. In order to define this quantity we first need to give some definitions that play an important role in the theory of curves.

Definitions. Let $\mathbf{x}$ be a regular smooth curve parametrized by arc length plus a constant (hence $|\mathbf{x}'| = 1$), assume that $\mathbf{x}$ has a continuous third derivative, and assume also that $\kappa \neq 0$ near the parameter value s_0. The *principal unit normal vector* at parameter value s is $\mathbf{N}(s) = |\mathbf{x}''(s)|^{-1}\,\mathbf{x}''(s)$. We have already encountered a special case of this vector in the study of curvatures and osculating planes, and if $\mathbf{T}(s) = \mathbf{x}'(s)$ denotes the unit tangent vector then we know that $\{\,\mathbf{T}(s),\,\mathbf{N}(s)\,\}$ is a set of perpendicular vectors with unit length (an *orthonormal* set).

If $\mathbf{x}$ is a space curve (*i.e.*, its image lies in 3-space), the *binormal* vector at parameter value s is defined to be $\mathbf{B}(s) = \mathbf{T}(s)\times\mathbf{N}(s)$. It then follows that $\{\,\mathbf{T}(s),\,\mathbf{N}(s),\,\mathbf{B}(s)\}$ is an orthonormal basis for $\mathbb{R}^3$, and it is called the *Frenet trihedron* (or frame) at parameter value s.

One can frequently define a Frenet trihedron at a parameter value s_0 even if the curvature vanishes at s_0, but there are examples where it is not possible to do so. In particular, consider the curve given by $\mathbf{x}(t) = \big(t,\,0,\,\exp(-1/t^2)\,\big)$ if $t > 0$ and $\mathbf{x}(t) = \big(t,\,\exp(-1/t^2)\,0\big)$ if $t > 0$. If we set $\mathbf{x}(t) = \mathbf{0}$, then $\mathbf{x}$ will be infinitely differentiable because for each $k \geq 0$ we have

$$\lim_{t\to 0}\ \frac{d^k}{dt^k}\,\exp(-1/t^2) \quad = \quad 0$$

(this is true by repeated application of L'Hospital's Rule) and in fact the curvature is also nonzero if $t \neq 0$ and $t^2 \neq 2/3$. Therefore one can define a principal unit normal vector $\mathbf{N}(t)$ when $t \neq 0$ but, say, $|t| < \frac{1}{2}$. However, if $t > 0$ this vector lies in the xz-plane while if $t < 0$ it lies in the xy-plane, and if one could define a continuous unit normal at $t = 0$ it would have to lie in both of these planes. Now the unit tangent at $t = 0$ is the unit vector $\mathbf{e}_1$, and there are no unit vectors that are perpendicular to $\mathbf{e}_1$ that lie in both the xy- and xz-planes. Therefore there is no way to define a continuous extension of $\mathbf{N}$ to all values of t. On the other hand, Problem 4.15 on pages 75–76 of Schaum's Outline Series on Differential Geometry provides a way to define principal normals in some situations when the curvature vanishes at a given parameter value.■

The following online notes contain further information on defining a parametrized family of moving orthonormal frames associated to a regular smooth curve:

http://ada.math.uga.edu/teaching/math4250/Html/Bishop.htm

One can retrieve the Frenet trihedron from an arbitrary regular smooth reparametrization with a continuous second derivative.

LEMMA. *In the setting above, suppose that we are given an arbitrary reparametrization with continuous second derivative, and let $s(t)$ denote the arc length function. Then the Frenet trihedron at parameter value t_0 is given by the unit vectors pointing in the same directions as $\mathbf{T}(t)$, $\mathbf{T}\,'(t)$, and their cross product. Furthermore, if one considers the reoriented curve $\mathbf{y}$ with parametrization $\mathbf{y}(t) = \mathbf{x}(-t)$, then the effect on the Frenet trihedron is that the first two unit vectors are sent to their negatives and the third remains unchanged.*

Proof. It follows immediately from the Chain Rule that the unit tangent $\mathbf{T}$ remains unchanged under a standard reparametrization with $s' > 0$. Furthermore, the derivation of the formulas for curvature under reparametrization show that $\mathbf{T}\,'(t)$ is a positive multiple of $\mathbf{x}''(s)$. this proves the assertion regarding the principal normals. Finally, if we are given two ordered sets of vectors $\{\,\mathbf{a},\,\mathbf{b}\,\}$ and $\{\,\mathbf{c},\,\mathbf{d}\,\}$ such that $\mathbf{c}$ and $\mathbf{d}$ are positive multiples of $\mathbf{a}$ and $\mathbf{b}$ respectively, then $\mathbf{c} \times \mathbf{d}$ is a positive multiple of $\mathbf{a} \times \mathbf{b}$, and this implies the statement regarding the binormals.

If one reverses orientations by the reparametrization $t \longmapsto -t$, then the Chain Rule implies that $\mathbf{T}$ and its derivative are sent to their negatives, and this proves the statement about the first two vectors in the trihedron. The statement about the third vector follows from these and the cross product identity $\mathbf{a} \times \mathbf{b} = (-\mathbf{a}) \times (-\mathbf{b})$.∎

We are finally ready to define torsion.

Definition. In the setting above the *torsion* of the curve is given by $\tau(s) = -\mathbf{B}\,'(s) \cdot \mathbf{N}(s)$.

The following alternate characterization of torsion is extremely useful in many contexts.

LEMMA. *The torsion of the curve is given by the formula $\mathbf{B}\,'(s) = -\tau(s)\,\mathbf{N}(s)$.*

Proof. If we can show that the left hand side is a multiple of $\mathbf{N}(s)$, then the formula will follow by taking dot products of both sides of the equation with $\mathbf{N}(s)$ (note that the dot product of the latter with itself is equal to 1). To show that the left hand side side is a multiple of $\mathbf{N}(s)$, it suffices to show that it is perpendicular to $\mathbf{T}(s)$ and $\mathbf{B}(s)$. The second of these follows because

$$0 \quad = \quad \frac{d}{ds}(1) \quad = \quad \frac{d}{ds}(\mathbf{B} \cdot \mathbf{B}) \quad = \quad 2\,\mathbf{B} \cdot \left(\frac{d\,\mathbf{B}}{ds}\right)$$

and the first follows because

$$\frac{d\mathbf{B}}{ds} \quad = \quad \frac{d}{ds}\,(\mathbf{T} \times \mathbf{N}) \quad = \quad (\kappa\,\mathbf{N} \times \mathbf{N}) \;+\; \left(\mathbf{T} \times \frac{d\,\mathbf{N}}{ds}\right) \quad = \quad \mathbf{T} \times \left(\frac{d\,\mathbf{N}}{ds}\right)$$

which implies that the left hand side is perpendicular to $\mathbf{T}$.∎

We had mentioned that the torsion of a curve is related to the rate at which a curve moves away from its osculating plane. Here is a more precise statement about the relationship:

PROPOSITION. *Let $\mathbf{x}$ be a regular smooth curve parametrized by arc length plus a constant (hence $|\mathbf{x}'| = 1$), assume that $\mathbf{x}$ has a continuous third derivative, and assume also that $\kappa(s_0) \neq 0$. Let Π be the osculating plane of $\mathbf{x}$ at parameter value s_0. Then the image of $\mathbf{x}$ is contained in Π for all s sufficiently close to s_0 if and only if the torsion vanishes for these parameter values.*

The following identity is useful for many purposes, and it will be used in the proof of the proposition: **LEMMA.** *If* $\mathbf{a}$, $\mathbf{b}$ *and* $\mathbf{c}$ *form an orthonormal basis of* $\mathbb{R}^3$, *then an arbitrary vector* $\mathbf{x}$ *in the latter can be written as*

$$\mathbf{x} \;\; = \;\; (\mathbf{x}\cdot\mathbf{a})\,\mathbf{a} \; + \; (\mathbf{x}\cdot\mathbf{a})\,\mathbf{b} \; + \; (\mathbf{x}\cdot\mathbf{c})\,\mathbf{c} \; .$$

This follows immediately by taking dot products of both sides with $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ respecitvely.■

Proof of the proposition. Suppose first that the curve is entirely contained in the osculating plane for s close to s_0. The osculating plane at s_0 is defined by the equation

$$[(\mathbf{y}-\mathbf{a}),\,\mathbf{T}_0,\,\mathbf{N}_0] \;\; = \;\; 0$$

where $\mathbf{a} = \mathbf{x}(s_0)$ and $\mathbf{T}_0$ and $\mathbf{N}_0$ represent the unit tangent and principal normal vectors at parameter value s_0. If we set $\mathbf{y} = \mathbf{x}(s)$ and simplify this expression, we see that the curve $\mathbf{x}$ satisfies the equation

$$\mathbf{x}(s)\cdot\mathbf{B}_0 \;\; = \;\; \mathbf{a}\cdot\mathbf{B}_0$$

where $\mathbf{B}_0 = \mathbf{T}_0 \times \mathbf{N}_0$. If we differentiate both sides with respect to s we obtain the equation $\mathbf{x}'(s)\cdot\mathbf{B}_0 = 0$. Differentiating once again we see that $\mathbf{x}''(s)\cdot\mathbf{B}_0 = 0$. Since $\mathbf{x}'(s) = \mathbf{T}(s)$ and $\mathbf{N}(s)$ is a positive multiple of $\mathbf{x}''(s)$ for s close to s_0 (specifically at least close enough so that $\kappa(s)$ is never zero), then $\mathbf{B}_0$ is perpendicular to both $\mathbf{T}(s)$ and $\mathbf{N}(s)$. Therefore the identity in the lemma implies that $\mathbf{B}_0 = (\mathbf{B}(s)\cdot\mathbf{B}_0)\,\mathbf{B}(s)$ for all s, so that $\mathbf{B}(s)$ must be a multiple of $\mathbf{B}(s)$ for all s; since both of these are unit vectors, it follows that $\mathbf{B}(s)$ must be equal to $\pm\,\mathbf{B}_0$. By continuity we must have that $\mathbf{B}(s) = \mathbf{B}_0$ for all s close to s_0 (Here are the details: Look at the function $\mathbf{B}(s)\cdot\mathbf{B}_0$ on some small interval containing s_0; its value is $\pm\,1$, and its value at s_0 is $+1$ — if its value somewhere else on the interval were -1, then by the Intermediate Value Theorem there would be someplace on the interval where its value would be zero, and we know this is impossible). Thus $\mathbf{B}(s)$ is constant, and by the preceding formulas this means that the torsion of the curve must be equal to zero.

Conversely, suppose that the torsion is identically zero. Then by alternate description of torsion in the lemma we know that $\mathbf{B}\,'(s) \equiv \mathbf{0}$, So that $\mathbf{B}(s) \equiv \mathbf{B}_0$. We then have the string of equations

$$0 \;\; = \;\; \mathbf{T}\cdot\mathbf{B}_0 \;\; = \;\; \mathbf{x}'\cdot\mathbf{B}_0 \;\; = \;\; \frac{d}{ds}\big(\mathbf{x}\cdot\mathbf{B}_0\big)$$

which in turn implies that $\mathbf{x}\cdot\mathbf{B_0}$ is a constant, and this constant must be $\mathbf{x}(s_0)\cdot\mathbf{B}_0$. Therefore the curve $\mathbf{x}$ lies entirely in the unique plane containing $\mathbf{x}(s_0)$ with normal direction $\mathbf{B}_0$.■

Examples. A **helix curve** given by parametric equations like $\mathbf{x}(t) = (\cos t, \sin t, t)$ is a simple example of a curve that is not planar. Curvature and torsion computations for curves of this type are given in `helix.pdf` and `helix2.pdf`.

Other planes associated to a curve

In addition to the osculating plane, there are two other associated planes through a point on the curve $\mathbf{x}$ at parameter value s_0 that are mentioned frequently in the literature. As above we assume that the curve is a regular smooth curve with a continuous third derivative i arc length parametrization, and nonzero curvature at parameter value s_0.

Definitions. In the above setting the *normal plane* is the unique plane containing $\mathbf{x}(s_0)$, $\mathbf{x}(s_0) + \mathbf{N}(s_0)$, and $\mathbf{x}(s_0) + \mathbf{B}(s_0)$, and the *rectifying plane* is the unique plane containing $\mathbf{x}(s_0)$, $\mathbf{x}(s_0) + \mathbf{T}(s_0)$, and $\mathbf{x}(s_0) + \mathbf{B}(s_0)$. These three mutually perpendicular planes meet at the point $\mathbf{x}(s_0)$ in the same way that the usual xy-, yz-, and xz-planes meet at the origin.

Oriented curvature for plane curves

For an arbitrary regular curve in 3-space one does not necessarily have normal directions when the curvature is zero, but for plane curves there is a unique normal direction up to sign. Specifically, if $\mathbf{x}$ is a regular smooth plane curve parametrized by arc length and $\mathbf{B}$ is a unit normal vector to a plane Π containing the image of $\mathbf{x}$, then one has an *associated oriented principal normal direction* at parameter value given by the cross product formula

$$\widehat{\mathbf{N}}\,(s) \quad = \quad \mathbf{B} \times \mathbf{x}'(s)$$

and by construction Π is the unique plane passing through $\mathbf{x}(s)$, $\mathbf{x}(s) + \mathbf{x}'(s)$, and $\mathbf{x}(s) = \widehat{\mathbf{N}}\,(s)$. There are two choices of $\mathbf{B}$ (the two unit normals for π are negatives of each other) and thus there are two choices for $\widehat{\mathbf{N}}\,(s)$ such that each is the negative of the other. One can then define a *signed curvature* associated to the oriented principal normal $\widehat{\mathbf{N}}$ given by the formula

$$k(s) \quad = \quad \left(\mathbf{x}''(s) \,\cdot\, \widehat{\mathbf{N}}\,(s)\right)$$

and since $\mathbf{x}''(s)$ is perpendicular to $\mathbf{x}'(s)$ and $\mathbf{B}$ this may be rewritten in the form

$$\mathbf{x}''(s) \quad = \quad k(s)\,\widehat{\mathbf{N}}\,(s)\ .$$

An obvious question is to ask what happens if $\kappa(s_0) = 0$ (which also equals $k(s)$ in this case) and the sign of $k(s)$ is negative for $s < s_0$ and positive for $s > s_0$. A basic example of this sort is given by the graph of $f(x) = x^3$ near $x = 0$, whose standard parametrization is given by $\mathbf{x}(t) = (t,\, t^3)$. In this situation the graph lies in the lower half plane $y < 0$ for $t < 0$ and in the in the upper half plane $y > 0$ for $t > 0$, and the curve switches from being concave upward for $t < 0$ to concave downward (generally called *convex* beyond first year calculus courses). The file `signed_curvature.pdf` looks at this example more closely, and in particular the computations in that file show that the signed curvature of the graph is positive if $t > 0$ and negative if $t < 0$. More generally, one usually says that f has a *point of inflection* in such cases.

The following result shows that more general plane curves behave similarly provided the curvature has a nonvanishing derivative:

PROPOSITION. *Let $\mathbf{x}$ be a regular plane smooth curve parametrized by arc length plus a constant (hence $|\mathbf{x}'| = 1$), assume that $\mathbf{x}$ has a continuous fourth derivative, let $\widehat{\mathbf{N}}$ define a family of oriented principal normals for $\mathbf{x}$, and assume that that $k(s_0) = 0$ but $k'(s_0) > 0$. Then $\mathbf{x}(s)$ is contained in the half plane*

$$\widehat{\mathbf{N}}\,(s_0) \cdot \big(\mathbf{y} - \mathbf{x}(s_0)\big) \quad < \quad 0$$

for s sufficiently close to s_0 satisfying $s < s_0$, and $\mathbf{x}(s)$ is contained in the half plane

$$\widehat{\mathbf{N}}\,(s_0) \cdot \big(\mathbf{y} - \mathbf{x}(s_0)\big) \quad > \quad 0$$

for s sufficiently close to s_0 satisfying $s > s_0$.

A similar result holds if $k'(s_0) < 0$, and the necessary modifications of the statement and proof for that case are left to the reader as an exercise.

Proof. To simplify the computations we shall choose coordinate systems such that $\mathbf{x}(s_0) = \mathbf{0}$ and the plane is the standard coordinate plane through the origin with chosed unit normal vector $\mathbf{e}_3$. It will also be convenient to denote the unit vector $\mathbf{x}'(s)$ by $\mathbf{T}(s)$. We shall need to work with a third order approximation to the curve, which means that we are going to need some information about $\mathbf{x}'''(s_0)$. Therefore the first step will be to establish the following formula:

$$k'(s_0) \quad = \quad \mathbf{x}'''(s_0) \ \cdot \ \widehat{\mathbf{N}}\,(s_0)$$

To see this, note that

$$k'(s) \quad = \quad \frac{d}{ds}\Big(\mathbf{x}''' \ \cdot \ \widehat{\mathbf{N}}\Big) \quad =$$

$$\Big(\mathbf{x}'''(s) \ \cdot \ \widehat{\mathbf{N}}\,(s)\Big) \ + \ \Big(\mathbf{x}''(s) \ \cdot \ \widehat{\mathbf{N}}'(s)\Big) \quad = \quad \Big(\mathbf{x}'''(s) \ \cdot \ \widehat{\mathbf{N}}\,(s)\Big) \ + \ \Big(\widehat{\mathbf{N}}\,(s) \ \cdot \ \widehat{\mathbf{N}}'(s)\Big)$$

and the second summand in the right hand expression vanishes because $|\,\widehat{\mathbf{N}}\,|^2$ is always equal to 1 (this is the same argument which implies that the unit tangent vector function is perpendicular to its derivative).

Turning to the proof of the main result, the preceding paragraph and earlier consideration show that the curve $\mathbf{x}$ is given near s_0 by the formula

$$\mathbf{x}(s) \quad = \quad (s - s_0)\,\mathbf{T}(s_0) \ + \ \frac{k(s)\,(s - s_0)^2}{2}\,\widehat{\mathbf{N}}\,(s_0) \ + \ \frac{(s - s_0)^3}{3!}\,\mathbf{x}'''(s_0) \ + \ (s - s_0)^4\,\theta(s)$$

where $\theta(s)$ is bounded for s suffieicntly close to zero. To simplify notation further we shall write $\Delta s = s - s_0$.

If we take the dot product of the preceding equation with $\widehat{\mathbf{N}}\,(s_0)$ we obtain the formula, in which $y(s)$ is the dot product of $\theta(s)$ and $\widehat{\mathbf{N}}\,(s_0)$, so that $y(s)$ is also bounded for s sufficiently close to s_0:

$$\Big(\mathbf{x}(s) \cdot \widehat{\mathbf{N}}\,(s_0)\Big) \quad = \quad \frac{k'(s_0)}{3!}\,(\Delta s)^3 + y(s)\,(\Delta s)^4$$

If s is nonzero but sufficiently close to zero then the sign of the right hand side is equal to the sign of Δs because

(i) the sign of the first term is equal to the sign of Δs,

(ii) if we let M be a positive upper bound for $|y(s)|$ and further restrict Δs so that

$$|\Delta s| \quad < \quad \frac{k'(s_0)}{6B}$$

then the absolute value of the second term in the dot product formula will be less than the absolute value of the first term.

It follows that the sign of the dot product

$$\Big(\mathbf{x}(s) \cdot \widehat{\mathbf{N}}\,(s_0)\Big)$$

is the same as the sign of the inital term

$$\frac{k'(s_0)}{3!}\,(\Delta s)^3$$

which in turn is equal to the sign of Δs. Since the dot product has the same sign as Δs for $s \neq 0$ and s sufficiently small, it follows that $\mathbf{x}(s)$ lies on the half plane defined by the inequality $\mathbf{y}\cdot\widehat{\mathbf{N}}\,(s_0) < 0$ if $s < s_0$ and $\mathbf{x}(s)$ lies on the half plane defined by the inequality $\mathbf{y}\cdot\widehat{\mathbf{N}}\,(s_0) > 0$ if $s > s_0$.■

In fact, the center of the osculating circle also switches sides when one goes from values of s that are less than s_0 to values of s that are greater than s_0. However, the proof takes considerably more work.

COMPLEMENT. *In the setting above, let $\mathbf{z}(s)$ denote the center of the osculating circle to $\mathbf{x}$ at parameter value at parameter value $s \neq s_0$ close to s_0 (this exists because the curvature is nonzero at such points). Then $\mathbf{z}(s)$ is contained in the half plane*

$$\widehat{\mathbf{N}}\,(s_0)\cdot\big(\mathbf{y}-\mathbf{x}(s_0)\big) < 0$$

for s sufficiently close to s_0 satisfying $s < s_0$, and $\mathbf{z}(s)$ is contained in the half plane

$$\widehat{\mathbf{N}}\,(s_0)\cdot\big(\mathbf{y}-\mathbf{x}(s_0)\big) > 0$$

for s sufficiently close to s_0 satisfying $s > s_0$.

Proof. We need to establish similar inequalities to those derived above if $\mathbf{x}(s)$ is replaced by $\mathbf{z}(s)$; note that the latter is not defined for parameter value s_0 because the formula involves the reciprocal of the curvature and the latter is zero at s_0.

The center of the osculating circle at parameter value $s \neq s_0$ was defined to be $\mathbf{x}+\kappa^{-1}\mathbf{N}$, where $\mathbf{N}$ is the ordinary principal normal; we claim that the latter is equal to $\mathbf{x}+k^{-1}\widehat{\mathbf{N}}$. By definition we have

$$\mathbf{x}'' \;=\; \kappa\,\mathbf{N} \;=\; k\,\widehat{\mathbf{N}}$$

and since $\kappa = \pm\,k$ is nonzero we know that $\kappa^2 = k^2$. Dividing the displayed equation by this common quantity yields the desired formula

$$\kappa^{-1}\mathbf{N} \;=\; k^{-1}\,\widehat{\mathbf{N}}\;.$$

Therefore the proof reduces to showing that the sign of

$$\left(\mathbf{x}(s) \;+\; \frac{1}{k(s)}\,\widehat{\mathbf{N}}\,(s)\right)\cdot\widehat{\mathbf{N}}\,(s_0)$$

is equal to the sign of Δs.

Using the formula for $\mathbf{x}(s)$ near s_0 that was derived before, we may rewrite the preceding expression as

$$h(s) \;=\; \frac{k'(s_0)}{3!}\,(\Delta s)^3 + y(s)\,(\Delta s)^4 \;+\; \frac{1}{k(s)}\,\widehat{\mathbf{N}}\,(s)\;\cdot\;\widehat{\mathbf{N}}\,(s_0)\;.$$

We need to show that $h(s)$ has the same sign as $k(s)$ and its reciprocal, and this will happen if

$$\ell(s) \;=\; h(s)-\frac{1}{k(s)} \;=\; \frac{k'(s_0)}{3!}\,(\Delta s)^3 + y(s)\,(\Delta s)^4 \;+\; \frac{1}{k(s)}\,\widehat{\mathbf{N}}\,(s)\;\cdot\;\Big(\widehat{\mathbf{N}}\,(s_0) \;-\; \widehat{\mathbf{N}}\,(s)\Big)$$

is bounded for $s \neq s_0$ sufficiently close to zero. To see, this, suppose that $|\ell(s)| \leq A$ for some $A > 0$. If we then choose $\delta > 0$ so that $|k(s)| < 1/A$ for for $|\Delta s| < \delta$ but $\Delta s \neq 0$, if will follow that

$$\Delta s \quad > \quad 0 \quad \Longrightarrow \quad h(s) \quad = \quad \frac{1}{k(s)} + \left(h(s) - \frac{1}{k(s)} \right) \quad > \quad A \; + \; (-A) \quad > \quad 0$$

and similarly with all inequalities reversed and A switched with $-A$ if $\Delta s < 0$.

In order to prove that $\ell(s)$ is bounded, it suffices to prove that each of the three summands is bounded for, say, $|\Delta s| \leq r$. The absolute value of the first is bounded by $k'(s_0)\, r^3/6$ and the absolute value of the second is bounded by $B\, r^4$ where B is a positive upper bound for $|y(s)|$. By the Cauchy-Schwarz inequality the absolute value of the third is bounded from above by

$$\frac{\left| \widehat{\mathbf{N}}\,(s) \;-\; \widehat{\mathbf{N}}\,(s_0) \right|}{|k(s)|}$$

and using the Mean Value Theorem we may estimate the numerator and denominator of this expression separately as follows:

(*i*) $\left| \widehat{\mathbf{N}}\,(s) - \widehat{\mathbf{N}}\,(s_0) \right| \leq P \cdot |\Delta s|$, where P is the maximum value of $|\,\widehat{\mathbf{N}}\,'\,|$ on $[s_0 - r,\ s_0 + r]$.

(*ii*) $k(s) = k'(S_1)\,\Delta s$ for some S_1 between s_0 and s, so if we choose r so small that $k' > 0$ on $[s_0 - r,\ s_0 + r]$, then $|k(s)| \geq Q\,\Delta s$, where $Q > 0$ is the minimum of k' on that interval.

It then follows that the quotient P/Q is an upper bound for the absolute value of the third term in the formula for $\ell(s)$, and therefore the latter itself is bounded. This completes the proof that $z(s)$ lies on the half plane described in the statement of the result.■

I.5 : Frenet-Serret Formulas

(Lipschutz, Chapter 5 and Appendix I)

In ordinary and multivariable calculus courses, a great deal of emphasis is often placed upon working specific examples, and as indicated in the discussion preceding Section I.1 of these notes there is a wide assortment of interesting curves that can be studied using the methods of the preceding sections. However, the course notes up to this point have not included the sorts of worked out examples that one sees in a calculus book. The Schaum's Outline Series book gives numerous examples, and the book by O'Neill does include a few examples, but there are far fewer than one might expect in comparison to standard calculus texts. We have reached a point in this course where the reasons for this difference should be explained.

We already touched upon one reason when we described computational techniques for finding the curvature of a curve. Even in simple cases, it can be extremely difficult — if not impossible — to write things out explicitly using pencil and paper along with the techniques and results that are taught in ordinary and multivariable calculus courses. For example, we noted that arc length reparametrizations often involve functions that ordinary calculus cannot handle in a straightforward manner. And the situation gets even worse when one considers certain types of curves that arise naturally in classical physics, most notably those arising when one attempts to describe the motions of a gravitational system involving three heavenly bodies. In these cases it is not even possible to give explicit formulas for the motion of the curves themselves, without even thinking about the added difficulty of describing quantities like curvature and torsion. During the past quarter century, spectacular advances in computer technology have provided powerful new tools for studying examples. A few comments on the use of computer graphics in differential geometry appear in O'NEILL. The following book is an excellent reference for further information on studying curves and surfaces using the software package *Mathematica*:

> A. Gray. Modern Differential Geometry of Curves and Surfaces. (Studies in Advanced Mathematics.) *CRC Press, Boca Raton, FL etc.*, 1993. ISBN: 0-8493-7872-9.

The emphasis in this course will be on *qualitative* aspects of the differential geometry of curves and surfaces in contrast to the *quantitative* emphasis that one sees in ordinary and multivariable calculus. In particular, we are interested in the following basic sort of question:

Reconstructing curves from partial data. *To what extent can one use geometric invariants of a curve such as curvature and torsion to retrieve the original curve?*

Both curvature and torsion are defined so that they do not change if one replaces a curve by its image under some rigid motion of $\mathbb{R}^2$ or $\mathbb{R}^3$, so clearly the best we can hope for is to retrieve a curve up to some transformation by a rigid motion. The main results of this section show that curvature and torsion suffice to recover the original curve in a wide range of "reasonable" cases.

The crucial input needed to prove such results comes from the *Frenet-Serret Formulas*, which describe the derivatives of the three fundamental unit vectors in the Frenet trihedron associated to a regular smooth curve.

FRENET–SERRET FORMULAS. *Let* $\mathbf{x}$ *be a regular smooth curve parametrized by arc length (hence* $|\mathbf{x}'| = 1$*), assume that* $\mathbf{x}$ *has a continuous third derivative, and assume also that* $\kappa(s_0) \neq 0$*. Let* $\mathbf{T}(s)$*,* $\mathbf{N}(s)$ *and* $\mathbf{B}(s)$ *be the tangent, principal normal and binormal vectors in the Frenet trihedron for* $\mathbf{x}$ *at parameter value* s_0*. Then the following equations describe the derivatives of the vectors in the Frenet trihedron:*

$$\begin{array}{rcrrrr} \mathbf{T}' & = & & & \kappa\,\mathbf{N} & \\ \mathbf{N}' & = & - & \kappa\,\mathbf{T} & & + \quad \tau\,\mathbf{B} \\ \mathbf{B}' & = & & & -\quad \tau\,\mathbf{N} & \end{array}$$

Proof. We have already noted that the first and third equations are direct consequences of the definition of curvature and torsion. To derive the second equation, we take the identity $\mathbf{N} = \mathbf{B} \times \mathbf{T}$ and differentiate it with respect to s:

$$\mathbf{N}'(s) \quad = \quad \mathbf{B}'(s) \times \mathbf{T}(s) \ + \ \mathbf{B}(s) \times \mathbf{T}'(s) \quad =$$

$$-\tau(s)\left(\mathbf{N}(s) \times \mathbf{T}(s)\right) \ + \ \kappa\left(\mathbf{B}(s) \times \mathbf{N}(s)\right)$$

Since $\mathbf{T}$, $\mathbf{N}$ and $\mathbf{B}$ are mutually perpendicular unit vectors such that $\mathbf{B} = \mathbf{T} \times \mathbf{N}$, as usual the "BAC–CAB" rule for threefold cross products implies that $\mathbf{N} \times \mathbf{T} = -\mathbf{B}$ and

$$\mathbf{B} \times \mathbf{N} \quad = \quad -\mathbf{N} \times \mathbf{B} \quad = \quad -\mathbf{N} \times (\mathbf{T} \times \mathbf{N}) \quad = \ = -\mathbf{T}\ .$$

If we make these substitions into the displayed equations we obtain the second of the Frenet-Serret Formulas.∎

The significance of the Frenet-Serret formulas is that they allow one to describe a curve in terms of its curvature and torsion in an essentially complete manner. Conversely, it follows that every pair of functions κ and τ with $\kappa(s) > 0$ for all s can be realized as the curvature and torsion functions for some curve. The Frenet-Serret formulas are the key to proving these results, but the proofs also require some facts about solutions to systems of linear differential equations, so a digression is needed to prove the necessary results about such systems.

The exponential function for matrices

The Frenet-Serret formulas can be written as a matrix differential equation $\mathbf{F}' = A(s)\,\mathbf{F}$, where $\mathbf{F}$ denotes the Frenet trihedron $(\,\mathbf{T},\mathbf{N},\mathbf{B}\,)$ and $A(s)$ is the skew-symmetric 3×3 matrix

$$\begin{pmatrix} 0 & \kappa(s) & 0 \\ -\kappa(s) & 0 & \tau(s) \\ 0 & -\tau(s) & 0 \end{pmatrix} .$$

There is an exponential function on square matrices which provides a concise way of describing the solutions to such systems of homogeneous linear differential equations, and it parallels the use of ordinary exponentials to solve simple differential equations of the form $y' = \lambda\, y$. For square matrices the exponential function can be defined by the same sort of infinite series used in calculus courses, but some work is needed in order to justify the construction of such an infinite sum. Therefore we begin with some material needed to prove that certain infinite sums of matrices can be defined in a mathematically sound manner and have reasonable properties.

Limits and infinite series of matrices

Limits of vector valued sequences in $\mathbb{R}^n$ can be defined and manipulated much like limits of scalar valued sequences, the key adjustment being that distances between real numbers that are expressed in the form $|s-t|$ are replaced by distances between vectors expressed in the form $|\mathbf{x}-\mathbf{y}|$. Similarly, one can talk about convergence of a vector valued infinite series $\sum_{n=0}^{\infty} \mathbf{v}_n$ in terms of the convergence of the sequence of partial sums $\mathbf{s}_n = \sum_{i=0}^{n} \mathbf{v}_k$. As in the case of ordinary infinite series, the best form of convergence is absolute convergence, which corresponds to the convergence of the real valued infinite series $\sum_{n=0}^{\infty} |\mathbf{v}_n|$ with nonnegative terms. A fundamental theorem states that a vector valued infinite series converges if the auxiliary series $\sum_{n=0}^{\infty} |\mathbf{v}_n|$ does, and there is a generalization of the standard M-test: *If* $|\mathbf{v}_n| \leq M_n$ *for all* n *where* $\sum_n M_n$ *converges, then* $\sum_n \mathbf{v}_n$ *also converges.*

We can view $m \times n$ matrices as mn-dimensional coordinate vectors, and we shall say that the *Euclidean magnitude* of a matrix is the usual length of the associated mn-dimensional vector.

In order to work with infinite series of matrices, we need some information about how the Euclidean magnitude behaves with respect to matrix products that is similar to the standard rule $|u\,v| = |u| \cdot |v|$ for absolute values. The following result provides the key estimate.

PRODUCT MAGNITUDE ESTIMATE. *Let A and B be matrices (not necessarily square) so that the product AB is defined, and for an arbitrary matrix C let $\|C\|$ be its Euclidean magnitude. Then $\|A\,B\| \leq \|A\| \cdot \|B\|$.*

Proof. It suffices to prove that the squares of the left and right hand sides are unequal in the same order. This is helpful because the squares of the Euclidean magnitudes are the sums of the squares of the matrix entries.

Given a matrix P let $\mathrm{Row}_i(P)$ and $\mathrm{Col}_j(P)$ denotes its $i*$th row and j^{th} column respectively. We then have

$$\|A\,B\|^2 \quad = \quad \sum_{i,j=1}^{n} \left(\mathrm{Row}_i(A)\cdot \mathrm{Col}_j(B)\right)^2$$

and applying the Schwarz inequality to each term in the sum we see that the latter is less than or equal to

$$\sum_{i,j=1}^{n} \left|\mathrm{Row}_i(A)\right|^2 \cdot \left|\mathrm{Col}_j(B)\right|^2 \quad = \quad \left(\sum_{i=1}^{n} \left|\mathrm{Row}_i(A)\right|^2\right) \cdot \left(\sum_{j=1}^{n} \left|\mathrm{Col}_j(B)\right|^2\right) .$$

But $\|A\|^2$ is equal to the first factor of this expression and $\|B\|^2$ is equal to the second.■

One consequence of this estimate is the following matrix version of a simple identity for sums of infinite series:

INFINITE SUM FACTORIZATION. *Let $\sum_{k=1}^{\infty} A_k$ be a convergent infinite series of $m\times n$ matrices with sum S, and let P and Q be $s\times m$ and $n\times t$ matrices respectively. Then $\sum_{k=1}^{\infty} P\,A_k$ and $\sum_{k=1}^{\infty} A_k\,Q$ converge to $P\,S$ and $S\,Q$ respectively.*

Proof. Let S_r be the r^{th} partial sum of the original series. Then $P\,S_r$ and $S_r\,Q$ are the corresponding partial sums for the other two series, and we need to show that these two matrices become arbitrarily close to $P\,S$ and $S\,Q$ if r is sufficiently large. By the hypothesis we know the analogous statement is true for the original infinite series.

Let $\varepsilon > 0$ be given, and let L be the maximum of $\|P\|+1$ and $\|Q\|+1$. Choose R so large that $\|S_r - S\| < \varepsilon/L$ if $r \geq R$. It then follows that

$$\|P\,S_r - P\,S\| \quad \leq \quad \|P\|\cdot|\,S_r - S\| \quad < \quad \varepsilon$$

and similarly we have

$$\|S_r\,Q - S\,Q\| \quad \leq \quad \|Q\|\cdot|\,S_r - S\| \quad < \quad \varepsilon$$

so that the limits of the partial sums have their predicted values.■

Power series of matrices

In order to work with power series of matrices having the form

$$\sum_{k=0}^{\infty} c_k\,A^k$$

for suitable coefficients c_k, we need the following consequence of the Product Magnitude Estimate:

POWER MAGNITUDE ESTIMATE. *If A is a square matrix, then for all integers $k \geq 1$ we have $\|A^k\| \leq \|A\|^k$.*

Proof. This is a tautology if $k=1$ so proceed by induction, assuming the result is true for $k-1 \geq 1$. Then $A^k = A\,A^{k-1}$ and therefore by the preceding result and the induction hypothesis we have

$$\|A^k\| \quad = \quad \|A\,A^{k-1}\| \quad \leq \quad \|A\|\cdot\|A^{k-1}\| \quad \leq \quad |A\|\cdot\|A\|^{k-1} \quad = \quad \|A\|^k\ .■$$

COROLLARY. *Suppose that we are given a sequence of scalars c_k for which*

$$\lim_{k \to \infty} \frac{|c_{k+1}|}{|c_k|} \quad = \quad L$$

and A is a nonzero square matrix such that $\|A\|^{-1} > L$. Then the infinite matrix power series

$$\sum_{k=0}^{\infty} c_k \, A^k$$

converges absolutely.

Proof. The argument is closely related to the proof of the ratio test for ordinary infinite series. Upper estimates for the Euclidean magnitudes of the terms are given by the inequalities

$$\|c_k \, A^k\| \quad \leq \quad |c_k| \cdot \|A\|^k$$

and the latter converges if $\|A\|^{-1} > L$ by the ratio test. But this means that the matrix power series converges absolutely.■

SPECIAL CASE. *If A is a square matrix, then the* **exponential series**

$$\exp(A) \quad = \quad \sum_{k=0}^{\infty} \frac{1}{k!} \, A^k$$

converges absolutely.■

Properties of matrix exponentials

It follows immediately that $\exp(0) = I$, and there is also a weak version of the usual law of exponents $e^{a+b} = e^a \, e^b$:

PRODUCT FORMULA. *If A and B are commuting matrices of the same size (i.e, $A\,B = B\,A$), then* $\exp(A + B) = \exp(A) \cdot \exp(B)$.

Idea of the proof. As for ordinary infinite series, one needs to do a little work in order to view the product of two infinite series sums as the sum of a third infinite series. Specifically, if one starts with convergent infinite series $\sum_k \; u_k$ and $\sum_k \; v_k$ with sums U and V, then one wants to say that $U\,V = \sum_k \; w_k$, where

$$w_k = \sum_{p+q=k} u_p \cdot v_q \; .$$

This turns out to be true if the original sequences are absolutely convergent, and one can use the same proof in the situation presented here because we know that A and B commute.■

It is important to note that the product formula does not necessarily hold if AB and BA are not equal.

COROLLARY. *For all square matrices A the exponential* $\exp(A)$ *is invertible and its inverse is* $\exp(-A)$.■

If the square matrix A is similar to a matrix B that has less complicated entries (for example, if A is similar to a diagonal matrix B), then the following result is often very helpful in understanding the behavior of $\exp(A)$.

SIMILARITY FORMULA. *Suppose that A and B are similar $n \times n$ matrices such that $B = P^{-1}\,A\,P$ for some invertible matrix P, then* $\exp(B) = P^{-1}\,\exp(A)\,P$.

Proof. By definition we have

$$P^{-1}\,\exp(A)\,P \quad = \quad P^{-1}\,\left(\sum_{k=0}^{\infty}\,\frac{1}{k!}\,A^k\right)\,P$$

and by the Infinite Sum Factorization formulas (proven above) and the identity $P^{-1}\,A^k\,P = (P^{-1}\,A\,P)^k$ we know that the right hand side is equal to

$$\sum_{k=0}^{\infty}\,\frac{1}{k!}\,P^{-1}\left(A^k\right)\,P \quad = \quad \sum_{k=0}^{\infty}\,\frac{1}{k!}\,\left(P^{-1}\,A\,P\right)^k \quad = \quad \sum_{k=0}^{\infty}\,\frac{1}{k!}\,B^k \quad = \quad \exp(B)$$

which is what we wanted to prove.■

Differentiating matrix valued functions

Differentiation of a matrix valued function of one real variable makes sense so long as the scalar valued functions corresponding to all the matrix entries are differentiable, and in this case one defines the derivative entry by entry. These derivatives have many familiar properties:

If C is a constant then $C' = 0$.

$(A + B)' = A' + B'$.

$(k\,A)' = k\,A' + k'\,A$.

$(A\,B)' = A'\,B + A\,B'$.

Perhaps the most noteworthy point is that **one must watch the order of multiplication in the last of these identities**.

Just as for ordinary power series, one has good term by term differentiation properties, and the proofs for ordinary infinite series go through with minimal changes:

TERMWISE DIFFERENTIATION. *Suppose that we have an infinite power series of $m \times n$ matrices $B(t) = \sum_{k=0}^{\infty} t^k\,B_k$ such that the radius of convergence for the auxiliary series $\beta(t) = \sum_{k=0}^{\infty} t^k\,\|B_k\|$ is at least r. Then the radius of convergence of $B(t)$ is at least r, and inside this radius of convergence we have $B'(t) = \sum_{k=0}^{\infty} t^k\,(k+1)\,B_{k+1}$.*■

If we apply this to the matrix exponential function $F(t) = \exp(tA)$ we obtain the equation

$$F'(t) \quad = \quad A\,\exp(tA) \quad = \quad \exp(tA)\,A\ .$$

All of this leads to the following result:

THEOREM. *For a given $n \times n$ matrix A and an $n \times 1$ column vector $\mathbf{b}$, there is a unique solution to the linear system of differential equations $X' = A\,X$ with initial condition $X(0) = \mathbf{b}$, and it is given by* $\exp(tA)\,\mathbf{b}$.

Proof. We first verify that the function described above is indeed a solution by applying the Leibniz rule. If $F(t) = \exp(tA)$, the latter says that the derivative of the function is given by the derivative of $F(t)\,\mathbf{b}$, which is equal to $F'(t)\,\mathbf{b}$, and by the discussion before the statement of the theorem this is equal to $A\,F(t)\,\mathbf{b}$. Also, the value at $t = 0$ is $\mathbf{b}$ because $\exp(0) = I$.

Conversely, suppose now that $X(t)$ solves the system of differential equation and has initial condition $X(0) = \mathbf{b}$. We proceed in analogy with the standard case where $n = 1$ and consider the product

$$W(t) \quad = \quad \exp(tA)^{-1}\,X(t) \quad = \quad exp(-tA)\,X(t)\ .$$

If we differentiate and apply the Leibniz Rule we obtain the following:

$$W'(t) \;=\; -A\,\exp(-tA)\,X(t) + \exp(-tA)\,X'(t) \;=\; -\exp(-tA)\,A\,X(t) + \exp(-tA)\,A\,X(t) \;=\; 0\,.$$

Therefore $W(t)$ is constant and equal to $W(0) = \mathbf{b}$. Left multiplication by $\exp(tA)$ then yields

$$\exp(tA)\,\mathbf{b} \quad = \quad \exp(tA)\,W(t) \quad = \quad \exp(tA)\,\exp(-tA)\,X(t) \quad = \quad I\,X(t) \quad = \quad X(t)\ .$$

This proves that the exponential solution is the only one with initial condition $X(0) = \mathbf{b}$.■

Differential equations with variable coefficients

The same methods also allow one to solve systems of differential equations of the form $X'(t) = A(t)\,X(t)$, where $X(t)$ is an $n \times p$ matrix and $A(t)$ is an $n \times n$ matrix whose coefficients are, say, continuous, or have continuous derivatives. The idea is very much the same one which is applied to solve ordinary first order linear differential equations, and here is the main result:

THEOREM. *For a given $n \times n$ matrix $A(t)$ of functions whose entries have continuous derivatives, and an $n \times p$ matrix vector C_0, there is a unique solution to the linear system of differential equations $X'(t) = A(t)\,X(t)$ with initial condition $X(0) = C_0$, and it is given by $\exp(\,P(t))\;C_0$, where $P'(t) = A(t)$ and $P(0) = 0$.*

The proof follows a standard pattern. One uses the Chain Rule and Leibniz Rule to show that $\exp(\,P(t))\;C_0$ solves the differential equation and has the right initial value at $t = 0$. Conversely, if $X(t)$ solves the differential equation let

$$Y(t) \quad = \quad \exp(\,-P(t))\;X(t)\ .$$

Then direct calculation shows that $Y'(t) = 0$ so that $Y(t) = C_0$ for all t. if we left multiply both sides of this equation by $\exp(\,P(t))$, we find that $X(t)$ has the form described previously.■

COROLLARY. (Uniqueness Theorem for curves with prescribed curvature and torsion). *Suppose that $\mathbf{x}$ and $\mathbf{y}$ are two regular smooth curves parametrized by arclength such that both have the same value at $t = 0$, both have the same Frenet-Serret trihedron at $t = 0$, and both have the same curvature and torsion functions (with nonzero curvature everywhere) such that both functions have continuous derivatives. Then $\mathbf{x}$ and $\mathbf{y}$ are identical.*

Proof. Let $(\,\mathbf{T_x}, \mathbf{N_x}, \mathbf{B_x}\,)$ and $(\,\mathbf{T_y}, \mathbf{N_y}, \mathbf{B_y}\,)$ denote the Frenet trihedra for $\mathbf{x}$ and $\mathbf{y}$ respectively. Then the uniqueness result for systems of linear equations implies that the trihedra for $\mathbf{x}$ and $\mathbf{y}$ are the same. It follows that

$$\mathbf{x}(s) - \mathbf{x}(0) \quad = \quad \int_0^s \mathbf{T_x}(u)\,du \quad = \quad \int_0^s \mathbf{T_y}(u)\,du \quad = \quad \mathbf{y}(s) - \mathbf{y}(0)$$

and therefore the initial conditions $\mathbf{y}(0) = \mathbf{x}(0)$ imply that $\mathbf{y}(s) = \mathbf{x}(s)$ for all s.■

It remains to prove that there are curves realizing arbitrary data of the form

(a) initial point on curve at $t = 0$,

(b) initial Frenet trihedron at $t = 0$,

(c) prescribed curvature and torsion functions,

such that the data satisfy the Frenet-Serret conditions. The existence result for solutions to systems of linear differential equations strongly suggests that this is true, but there are a few things that we need to check.

Skew-symmetric matrices, orthogonal matrices and exponentials

The Frenet-Serret formulas are a system of differential equations as above such that $n = p = 3$ and $A(t)$ is a skew-symmetric matrix. The Frenet-Serret system of differential equations has a solution with given initial trihedron, but we need to know that the solution

$$X(s) \quad = \quad (\,\mathbf{T}(s), \mathbf{N}(s), \mathbf{B}(s)\,)$$

has orthonormal columns and the 3×3 matrix on the right hand side has determinant equal to $+1$ (since $\mathbf{B} = \mathbf{T} \times \mathbf{N}$).

Recall that an $n \times n$ matrix P is said to be an *orthogonal matrix* if it satisfies the condition $P^*P = PP^* = I$ where P^* denotes the transpose of P. By the definitions of matrix multiplication and dot products, the entries of the product P^*P are the dot products of various pairs of columns in P, so that a matrix with orthonormal columns satisfies $P^*P = I$; this condition implies that P has maximum rank n and hence is invertible, which in turn implies that $PP^* = I$. In other words, the matrix P has orthonormal columns if and only if P is orthogonal.

This relates to solutions of the linear system $X'(s) = A(s)\,X(s)$ as follows: We need to show that if $A(s)$ is skew-symmetric and the initial condition $X(0)$ is an orthogonal matrix with determinant $+1$, then each $X(s)$ is also orthogonal with determinant $+1$, and the following result is basically what we need in order to verify this.

The following result gives us everything we need:

THEOREM. *If A is a skew-symmetric $n \times n$ matrix, then* $\exp(A)$ *is orthogonal.*

Before beginning the proof, we note that if $P(t)$ is an arbitrary matrix valued function of one variable, then we have the elementary identity

$$(P^*)' \quad = \quad (P')^*$$

and standard identities for matrix transposition yield the identity

$$\exp(A^*) \quad = \quad \exp(A)^* \ .$$

Proof of the theorem. Suppose that $A^* = -A$. Then A^* and A obviously commute, and therefore we have

$$I \quad = \quad \exp(0 = A + A^*) \quad = \quad \exp(A)\,\exp(A^*) \quad = \quad \exp(A)\,\exp(A)^*$$

and similarly $I = \exp(A^) * \exp(A)$, so that $\exp(A)$ is an orthogonal matrix.■

We can summarize everything into the **Fundamental Theorem of Space Curve Theory:**

Given smooth functions κ *and* τ *such that the first is always positive, an initial vector* $\mathbf{x}_0$ *and an orthonormal set of vectors* $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ *such that* $\mathbf{a} \times \mathbf{b} = \mathbf{c}$, *then there is a unique (sufficiently differentiable) curve* $\mathbf{x}$ *such that the tangent vectors to* $\mathbf{x}$ *at all point have unit length, the Frenet trihedron of* $\mathbf{x}$ *at* 0 *is given by the standard unit vectors*

$$\Big(\mathbf{T}(0),\, \mathbf{N}(0),\, \mathbf{B}(0)\Big) \quad = \quad (\mathbf{a},\, \mathbf{b},\, \mathbf{c})$$

and the curvature and torsion functions are respectively given by κ *and* τ.■

In particular, this result implies that *space curves are completely determined by their curvature and torsion functions together with the Frenet trihedron at some initial value.*

Proof. The solution of the system of differntial equations for the Frenet trihedron has the form $\exp(P(s))\, F_0$ where $P(s)$ is skew-symmetric and the initial value F_0 is an orthogonal matrix with determinant $+1$. Standard matrix algebra identities imply that a product of two orthogonal matrices is orthogonal; since $\exp(P(s))$ is orthogonal by the previous theorem, it follows that the solution to the Frenet-Serret system is always an orthogonal matrix. Also, the condition $P^*P = I$ and standard determinant identities imply that the determinant of an orthogonal matrix is always ± 1, and hence this is true for the determinant of the solution $\exp(P(s))\, F_0$. Now the determinant of the latter is clearly a continuous function of s and since it is always either $+1$ or -1 the value of the determinant is a constant function of s. If $s = 0$ then $P(0)$ is the zero matrix by our construction of the explicit solution to the system of differential equations, so that $\exp(P(0)) = I$ and this constant value is $\det F_0 = +1$. Therefore the columns in the solution to the system of differential equations satisfy the orthonormality and right hand conditions required for a Frenet trihedron.

To complete the proof we need to show that the solution

$$\big(\,\mathbf{T}(s), \mathbf{N}(s), \mathbf{B}(s)\,\big)$$

actually corresponds to the Frenet-Serret data for some curve, but we can do this by defining a curve via the formula

$$\mathbf{x}(s) \quad = \quad \mathbf{x}(0) \; + \; \int_0^s \; \mathbf{T}(u)\, du \; .$$

It is necessary to check that the solution does give the Frenet trihedron for this curve, but doing so is a fairly straightforward exercise which is left to the reader (for example, the unit tangent vector function is equal to $\mathbf{T}$ by construction).■

The following special case of the Fundamental Theorem is a companion to our earlier characterization of lines as curves whose curvature is identically zero:

CHARACTERIZATION OF CIRCULAR ARCS. *Let* $\mathbf{x}$ *be a curve satisfying the conditions in the statement of the Frenet-Serret Formulas. Then* $\mathbf{x}$ *is a circular arc if and only if the curvature is a positive constant and the torsion is identically zero.*

This follows immediately because we can always find a circular arc with given initial value $\mathbf{x}_0$, initial Frenet trihedron $\big(\mathbf{T}_0,\, \mathbf{N}_0,\, \mathbf{B}_0\,\big)$ and constant curvature $\kappa > 0$ (and also of course with vanishing torsion); in fact, the equations for an osculating circle provide an explicit construction.■

Since plane curves may be viewed as space curves whose third coordinates are zero (and whose torsion functions are zero), the Fundamental Theorem of Local Curve Theory also applies to plane curves, and in fact the Fundamental Theorem amounts to saying that there is a unique curve with a given (nonzero) curvature function κ, initial value $\mathbf{x}_0$ and initial unit tangent vector $\mathbf{T}_0$; in this case the principal normal $\mathbf{N}_0$ is completely determined by the. perpendicularity condition and the Frenet-Serret Formulas.

In fact, there is actually a stronger version of the Fundamental Theorem in the planar case. In order to state and prove the Fundamental Theorem for space curves we needed to assume the curvature was positive so that the principal normal $\mathbf{N}$ could be defined. We have already noted that one can define $\mathbf{N}$ for plane curves even if the curvature is equal to zero. Geometrically, a standard way of doing this is to rotate the unit tangent $\mathbf{T}$ in the counterclockwise direction through an angle of $\pi/2$; in terms of equations this means that $\mathbf{N} = J(\mathbf{T})$, where J is the linear transformation

$$J(x,\, y) \quad = \quad (y,\, -x) \ .$$

As noted in the previous section, if $\mathbf{x}$ is a regular smooth curve in $\mathbb{R}^2$ parametrized by arc length plus a constant, this means that if we define an associated *signed curvature* by the formula

$$k(s) \quad = \quad \mathbf{x}''(s) \cdot \mathbf{N}(s) \quad = \quad \mathbf{x}''(s) \cdot [J(\mathbf{T})]\,(s)$$

then $|k(s)| = \kappa(s)$.

For the sake of completeness, we shall formally state and prove the modified version of the Frenet-Serret Formulas that holds in the 2-dimensional setting with $\mathbf{N}$ defined as above.

PLANAR FRENET–SERRET FORMULAS. *Let $\mathbf{x}$ be a regular smooth curve parametrized by arc length (hence $|\mathbf{x}'| = 1$), assume that $\mathbf{x}$ has a continuous third derivative. Let $\mathbf{T}(s)$ and $\mathbf{N}(s)$ and be the tangent and principal normal vectors for $\mathbf{x}$ at parameter value s_0. Then the following equations describe the derivatives of $\mathbf{T}$ and $\mathbf{N}$:*

$$\begin{array}{rcrl} \mathbf{T}' & = & & k\,\mathbf{N} \\ \mathbf{N}' & = & -\ k\,\mathbf{T} & \end{array}$$

Proof. By definition the first equation is a direct consequence of the definition of signed curvature. To derive the second equation, we take the identity $\mathbf{N}(s) = J(\,\mathbf{T}(s)\,)$ and differentiate it with respect to s, obtaining

$$\mathbf{N}'(s) \quad = \quad J(\,\mathbf{T}'(s)\,) \quad = \quad J\big(\,k(s)\,\mathbf{N}(s)\,\big) \quad = \quad k(s)\,J\Big(\,J\,(\,\mathbf{T}(s)\,)\,\Big) \quad =$$

$$k(s)\,J^2(\,\mathbf{T}(s)\,) \quad = \quad -k(s)\,\mathbf{T}(s)$$

where the last equation follows because $J^2 = -I$.■

One can use the notion of signed curvature to state and prove the following version of the fundamental theorem for plane curves:

FUNDAMENTAL THEOREM OF LOCAL PLANE CURVE THEORY. *Given a sufficiently differentiable function κ on some interval $(-c,\, c)$, an initial vector $\mathbf{x}_0$ and an orthonormal* *set of vectors* $(\mathbf{a},\,\mathbf{b})$ *such that* $\mathbf{b} = J(\mathbf{a})$*, then there is a sufficiently differentiable curve* $\mathbf{x}$ *such that* $\mathbf{x}(0) = \mathbf{x}_0$*, the tangent vectors to* $\mathbf{x}$ *at all point have unit length, the tangent-normal pair of* $\mathbf{x}$ *at at* 0 *is given by the standard unit vectors*

$$\Big(\mathbf{T}(0),\,\mathbf{N}(0)\Big) \quad = \quad \big(\mathbf{a},\,\mathbf{b}\,\big)$$

and the curvature function is given by κ*.*■

The proof of this result is a fairly straightforward modification of the argument for space curves and will not be worked out explicitly for that reason.

Local canonical forms

One application of the Frenet-Serret formulas is a description of a strong third order approximation to a curve in terms of curvature and torsion.

PROPOSITION. *Let* $\mathbf{x}$ *be a regular smooth curve parametrized by arc length plus a constant (hence* $|\mathbf{x}'| = 1$*) such that* $\mathbf{x}$ *has a continuous fourth derivative and* $\kappa(0) \neq 0$*, and let* $\{\,\mathbf{T},\,\mathbf{N},\,\mathbf{B}\,\}$ *be the Frenet trihedron at parameter value* $s = 0$*. Then a strong third order approximation to* $\mathbf{x}$ *is given by*

$$\mathbf{x}(0) \; + \; \left(s - \frac{s^2\kappa^2}{3!}\right)\mathbf{T} \; + \; \left(\frac{s^2\kappa}{2} - \frac{s^3\kappa'}{3!}\right)\mathbf{N} \; + \; \frac{s^3\kappa\tau}{3!}\mathbf{B}\;.$$

Proof. We already know that $\mathbf{x}'(0) = \mathbf{T}$ and $\mathbf{x}''(0) = \kappa\,\mathbf{N}$. It suffices to compute $\mathbf{x}'''(0)$, and the latter is given by

$$\big(\kappa\,\mathbf{N}\big)' \quad = \quad \kappa'\,\mathbf{N} \; + \; \kappa\,\mathbf{N}' \quad = \quad \kappa'\,\mathbf{N} \; - \; \kappa^2\,\mathbf{T} \; + \; \kappa\,\tau\mathbf{B}$$

where the last is derived using the Frenet-Serret Formulas.■

Here are two significant applications of the canonical form for the strong third order approximation. By the basic assumptions for the Frenet-Serret Formulas we have $\kappa > 0$.

APPLICATION 1. *In the setting above, if* $\tau(0) < 0$ *then the point* $\mathbf{x}(s)$ *lies on the side of the osculating plane defined by the inequality* $(\mathbf{y} - \mathbf{x}(0)) \cdot \mathbf{B} < 0$*, when* $s > 0$ *and* s *is sufficiently close to* 0*, and* $\mathbf{x}(s)$ *lies on the side of the osculating plane defined by the inequality* $(\mathbf{y} - \mathbf{x}(0)) \cdot \mathbf{B} > 0$ *when* $s < 0$ *and* s *is sufficiently close to* 0*. Similarly, if* $\tau(0) > 0$ *then the point* $\mathbf{x}(s)$ *lies on the side of the osculating plane defined by the inequality* $(\mathbf{y} - \mathbf{x}(0)) \cdot \mathbf{B} > 0$ *when* $s < 0$*, and* $\mathbf{x}(s)$ *lies on the side of the osculating plane defined by the inequality* $(\mathbf{y} - \mathbf{x}(0)) \cdot \mathbf{B} < 0$ *when* $s > 0$*.*■

Derivation. We shall only do the case where $\tau > 0$ and $s > 0$. The arguments in the other cases are basically the same, the main difference being that certain signs and inequality directions must be changed.

Let $g(s) = (\,\mathbf{x}(s) - \mathbf{x}(0)\,) \cdot \mathbf{B}$; then the orthonormality of the Frenet trihedron $\{\,\mathbf{T},\,\mathbf{N},\,\mathbf{B}\,\}$ and the canonical form yield the equation

$$g(s) \quad = \quad +\frac{s^3\kappa\tau}{3!} \; + \; \theta(s)$$

where $|\theta(s)| \leq |s|^4 \cdot M$ for some positive constant M. It follows that if $|s|$ is small and $s > 0$ then we have

$$g(s) \quad \geq \quad +\frac{s^3\kappa\tau}{3!} \; + \; M \cdot s^4$$

and the right hand side (hence also $g(s)$) is negative provided

$$s \quad < \quad \frac{\kappa|\tau|}{3!\,M} \ .\blacksquare$$

APPLICATION 2. *In the setting above, if $\kappa' \neq 0$ and $s \neq 0$ is sufficiently close to zero then $\mathbf{x}(s)$ lies on the side of the rectifying plane defined by the inequality*

$$\big(\mathbf{y} - \mathbf{x}(0)\big) \cdot \mathbf{N} \ < \ 0 \ .\blacksquare$$

Derivation. Let $g(s) = (\,\mathbf{x}(s) - \mathbf{x}(0)\,) \cdot \mathbf{N}$; then the canonical form implies an equation

$$g(s) \quad = \quad -\left(\frac{s^2\kappa}{2} \ + \ -\frac{s^3\kappa'}{3!}\right) \ + \ \theta(s)$$

where $|\theta(s)| \leq |s|^4 \cdot M$ for some positive constant M. We might as well assume that $M \geq 1$. It follows that if $|s|$ is small and nonzero then we have

$$|g(s)| \quad \geq \quad \left(\frac{s^2\kappa}{2} \ - \ \frac{|s|^3|\kappa'|}{3!}\right) \ - \ M \cdot |s|^4$$

and the right hand side is positive provided

$$|s| \quad < \quad \min\left(\frac{\kappa}{\kappa'} \ , \ \frac{\sqrt{\kappa}}{2M}\right) \ .$$

It follows that $g(s)$ is nonzero (and in fact negative) under the same conditions.$\blacksquare$

Regular smooth curves in hyperspace

During the nineteenth century mathematicians and physicists encountered numerous questions that had natural interpretations in terms of spaces of dimension greater than three (incidentally, in physics this began long before the viewing of the universe as a 4-dimensional space-time in relativity theory). In particular, coordinate geometry gave a powerful means of dealing with such objects by analogy. For example, Euclidean n-space for and arbitrary finite n is given by the vector space $\mathbb{R}^n$, lines, planes, and various sorts of hyperplanes can be defined and studied by algebraic methods (although geometric intuition often plays a key role in formulating, proving, and interpreting results!), and distances and angles can be defined using a simple generalization of the standard dot product. Furthermore, objects like a 4-dimensional hypercube or a 3-dimensional hypersphere can be described using familiar sorts of equations. For example, a typical hypercube is given by all points $\mathbf{x} = (x_1, x_2, x_3, x_4)$ such that $0 \leq x_i \leq 1$ for all i, and a typical hypersphere is given by all points $\mathbf{x}$ such that

$$|\mathbf{x}|^2 \quad = \quad x_1^2 \ + \ x_2^2 \ + \ x_3^2 \ + \ x_4^2 \quad = \quad 1 \ .$$

A full investigation of differential geometry in Euclidean spaces of dimension ≥ 4 is beyond the scope of this course, but some comments about the differential geometry of curves in 4-space seem worth mentioning.

One can define regular smooth curves, arc length and curvature for parametrized 4-dimensional curves exactly as for curves in 3-dimensional space. In fact, there are generalizations of the Frenet-Serret formula and the Fundamental Theorem of Local Curve Theory. One complicating factor is that the 3-dimensional cross product does not generalize to higher dimensions in a particularly neat fashion, but one can develop algebraic techniques to overcome this obstacle. In any case, in four dimensions if a sufficiently differentiable regular smooth curve $\mathbf{x}$ is parametrized by arc length plus a constant and has nonzero curvature and a nonzero secondary curvature (which is similar to the torsion of a curve in 3-space), then for each parameter value s there is an ordered orthonormal set of vectors $\mathbf{F}_i(s)$, where $1 \leq i \leq 4$, such that $\mathbf{F}_1$ is the unit tangent vector and the sequence of vector valued functions (the *Frenet frame* for the curve) satisfies the following system of differential equations, where κ_1 is curvature, κ_2 is positive valued, and the functions κ_1, κ_2, κ_3, all have sufficiently many derivatives:

$$\begin{array}{lclcccccc}
\mathbf{F}_1' & = & & & \kappa_1\,\mathbf{F}_2 & & & & \\
\mathbf{F}_2' & = & - & \kappa_1\,\mathbf{F}_1 & & + & \kappa_2\,\mathbf{F}_3 & & \\
\mathbf{F}_3' & = & & & -\kappa_2\,\mathbf{F}_2 & & & + & \kappa_3\,\mathbf{F}_4 \\
\mathbf{F}_4' & = & & & & & -\kappa_3\,\mathbf{F}_3 & &
\end{array}$$

The Fundamental Theorem of Local Curve Theory in 4-dimensional space states that locally there is a unique curve with prescribed higher curvature functions $\kappa_1 > 0$, $\kappa_2 > 0$ and κ_3, prescribed initial value $\mathbf{x}(s_0)$, and whose Frenet orthonormal frame satisfies $\mathbf{F}_i(s_0) = \mathbf{v}_i$ for some orthonormal basis $\{\,\mathbf{v}_1,\,\mathbf{v}_2,\,\mathbf{v}_3,\,\mathbf{v}_4\,\}$. An online description and derivation of such formulas in arbitrary dimensions is available at the sites

`http://www.math.technion.ac.il/~rbrooks/dgeo1.7.ps`

`http://en.wikipedia.org/wiki/Differential_geometry_of_curves`

and a discussion of such formulas in complete generality (*i.e.*, appropriate for a graduate level course) appears on page 74 of HICKS.

II. Topics from Geometry and Multivariable Calculus

This unit covers three topics involving background material. The first is a discussion of differential forms. These objects play a major role in O'NEILL's treatment of the subject, and we shall explain how one can pass back and forth between the classical vector formulations of concepts in differential geometry and their restatements in terms of the more modern (and ultimately more convenient) language of differential forms. Each approach appears frequently in the literature of the subject, so an understanding of their relationship is always useful and sometimes absolutely necessary. [*Note:* This material will note be covered in Winter2012.]

The second objective of this unit is to discuss some points regarding vector valued functions of several variables, and especially those which will be needed for studying surfaces in Units III and IV. One goal is to give concise and useful principles for working with such functions that closely resemble well known results in elementary calculus (*e.g.*, the linear approximation of functions near a point using derivatives, the Chain Rule, differentiability criteria for inverse functions, change of variables formulas in multiple integration). Finally, we shall use vector valued functions of several variables to give an analytic definition of congruence for geometric figures, and we shall combine this with the Frenet-Serret Formulas from Unit I to prove that two well behaved differentiable curves are congruent if and only if their curvature and torsion functions are equal.

II.1 : Differential forms

For the Winter 2012 course, only the subsections on multiple integration and connectedness will be covered.

During the 20^{th} century mathematicians and physicists discovered that many advanced topics in differential geometry could be handled more efficiently, and in greater generality, if certain concepts were reformulated from vector terminology into slightly different notation. The central objects in this setting are called *differential forms* or *exterior forms*. Among other things, differential forms provide answers to many cases of the following basic question:

Given a geometrical formula involving cross products in $\mathbb{R}^3$, *how can one generalize it to higher dimensions?*

A detailed answer to this question in terms of differential forms is beyond the scope of this course. However, O'NEILL works with differential forms frequently (but not exclusively), so it is worthwhile to explain how one can pass between the language of vectors and differential forms. One basic use of differential forms in differential geometry appears in Section 2.8 of O'NEILL, where an abstract analog of the Frenet-Serret Formulas is described. Chapters 6 and 7 of O'NEILL discuss some other basic aspects of classical differential geometry using differential forms.

BACKGROUND ON MULTIPLE INTEGRATION. The definition of differential forms is motivated by concepts involving double and triple integrals, so it will be necessary to discuss such objects here. More precisely, we shall need material from a typical multivariable calculus course or sequence through the main theorems from vector analysis. Files describing the background material (with references to standard texts used in the Department's courses) are included in the course directory under the names `background2.pdf`. Here are some further online references for background material:

`http://tutorial.math.lamar.edu/AllBrowsers/2415/DoubleIntegrals.asp`

`http://www.math.hmc.edu/calculus/tutorials/multipleintegration/`

`http://ndp.jct.ac.il/tutorials/Infitut2/node38.html`

`http://math.etsu.edu/MultiCalc/Chap4/intro.htm`

`http://www.maths.abdn.ac.uk/ igc/tch/ma2001/notes/node74.html`

`http://www.maths.soton.ac.uk/ cjh/ma156/handouts/integration.pdf`

`http://en.wikipedia.org/wiki/Multiple_integral`

Topics from multiple integration will also figure in a few subsequent sections, including the discussion of the Change of Variables Formula in Section II.3 and the remarks on surface area in Section III.5.

The basic objects

Everything can be done in $\mathbb{R}^n$ for all positive integers n, but we shall only need the cases where $n = 2$ or 3 in this course, so at some points our statements and derivations may only apply for these values of n.

Suppose that U is an open subset of $\mathbb{R}^n$,where $n = 2$ or 3. If $0 < p \leq n$, then a **differential** p**-form** may be described as follows.

The case $p = 1$. A 1-form is basically an integrand for line integrals over curves in U. Specifically, it has the form $\sum_i \; f_i \, dx_i$, where $1 \leq i \leq n$ and each f_i is a function on U with continuous partial derivatives.

The case $p = 2$. If $n = 2$, then a 2-form is basically an integrand for double integrals over subsets of U. Specifically, it has the form $f(x, y) \, dx \, dy$, where f has continuous partial derivatives. If $n = 3$, then a 2-form is basically an integrand for certain surface integrals over subsets of U (more precisely, flux integrals of vector fields taken over oriented surfaces). Specifically, these integrands have the form

$$P \, dy \, dz \; + \; Q \, dz \, dx \; + \; R \, dx \, dy$$

where P, Q, R are functions with continuous partial derivatives. For technical reasons that need not be discussed at this point, one inserts a wedge sign $\wedge$ between the second and third factors, so that a monomial form is written $H \, du \wedge dv$.

The case $p = 3$. This case only arises when $n = 3$, where a 3-form is basically an integrand for triple integrals over subsets of U. Specifically, it has the form $f(x, y, z) \, dx \, dy \, dz$, where f has continuous partial derivatives. As in the case $p = 2$, one interpolates wedges between the differential symbols dx, dy and dz so that the form is written $f(x, y, z) \, dx \wedge dy \wedge dz$.

Comparisons with vector fields

There is an obvious 1–1 correspondence between 1-forms and smooth vector fields, which we may view as vector valued functions $\mathbf{F}$ from U to $\mathbb{R}^n$ such that each coordinate function has continuous partial derivatives. Specifically, if the coordinates of $\mathbf{F}$ are $(P_1, ..., P_n)$, then $\mathbf{F}$ corresponds to the 1-form

$$\omega_{\mathbf{F}} \quad = \quad P_1\, dx_1 \ + \ \cdots \ + \ P_n\, dx_n$$

and conversely the right hand side determines a smooth vector field whose coordinates are the coefficients of the differential symbols dx_i.

Of course, it is natural to ask why one might wish to make such a looking change of notation. In particular, there should be some substantive advantage in doing so. One reason involves two basic themes in multivariable calculus: (1) The gradient of a function. (2) Change of variables formulas (*e.g.*, among rectangular, polar, cylindrical or spherical coordinates). We shall think of a change of variables as a generalization of the standard polar coordinate maps:

$$x \quad = \quad r \cos\theta \ , \qquad y \quad = \quad r \sin\theta$$

This takes open sets in the $r\,\theta$ plane to open sets in the $x\,y$-plane. Comparing the formulas for a function's gradient in two such coordinate systems can be extremely awkward. However, if we look at the *exterior derivative*

$$df \quad = \quad \sum_i \frac{\partial f}{\partial x_i}\, dx_i$$

rather than the gradient, then one obtains a much more tractable change of variables formula:

$$\frac{\partial f}{\partial x}\, dx \ + \ \frac{\partial f}{\partial y}\, dy \quad \longleftrightarrow \quad \frac{\partial f}{\partial r}\, dr \ + \ \frac{\partial f}{\partial \theta}\, d\theta$$

If $n = 3$, there is a different but related 1–1 correspondence between 2-forms and vector fields, in this case sending a vector field $\mathbf{F}$ with coordinate functions P, Q, R to the type of expression displayed above.

$$P\, dy \wedge dz \ + \ Q\, dz \wedge dx \ + \ R\, dx \wedge dy$$

The ∇ operator(s) and differential forms

The exterior derivative of a function is one case of a general construction of *exterior derivatives* on differential forms, which sends every p-form ω to a $(p+1)$-form $d\omega$; this can be extended to all nonnegative integers by agreeing that a 0-form is just a function and a p-form is zero if $p > n$. The formal definition is a bit complicated, but for our purposes it suffices to know that exterior differentiation is completely determined by the previous construction for df and following simple properties:

(1) For all forms ω we have $d(d\omega) = 0$.

(2) For all p forms ω and λ we have $d(\omega + \lambda) = d\omega + d\lambda$.

(3) For all p-forms ω and pure differential 1-forms dx_i we have $d(\omega \wedge dx_i) = d\omega \wedge dx_i$.

(4) For all pure differential 1-forms dx_i and dx_j we have $dx_i \wedge dx_j \ = \ -\, dx_j \wedge dx_i$ (hence it vanishes if $i = j$).

Verification of these for $n = 2$ or 3 reduce to a sequence of routine computations.

When one passes from the vector fields or scalar valued functions to differential forms, the ∇ operator(s) passes to exterior derivatives. Here is a formal statement of this correspondence.

THEOREM. *Let p and n be as above. The the following conclusions hold:*

(i) *Suppose that $p = 1$ and $n = 2$, and $\mathbf{F}$ is the vector field with coordinate functions (P, Q). If $\omega_{\mathbf{F}}$ is the differential 1-form corresponding to $\mathbf{F}$, then*

$$d\omega_{\mathbf{F}} \quad = \quad \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx\, dy \ .$$

(ii) *Suppose that $p = 1$ and $n = 3$, and $\mathbf{F}$ is the vector field with coordinate functions (P, Q). If $\omega_{\mathbf{F}}$ is the differential 1-form corresponding to $\mathbf{F}$, then*

$$d\omega_{\mathbf{F}} \quad = \quad \Omega_{\mathbf{G}}$$

where Ω_G denotes the 2-form corresponding to $\mathbf{G}$ and $\mathbf{G} = \nabla \times \mathbf{F}$ is the curl of $\mathbf{F}$.

(iii) *Suppose that $p = 2$ and $n = 3$, and $\mathbf{F}$ is the vector field with coordinate functions (P, Q, R). If $\Omega_{\mathbf{F}}$ is the differential 2-form corresponding to $\mathbf{F}$, then*

$$d\Omega_{\mathbf{F}} \quad = \quad (\nabla \cdot \mathbf{F})\, dx\, dy\, dz$$

where $\nabla \cdot \mathbf{F}$ denotes the divergence of $\mathbf{F}$.

Verifying each of these is a routine computational exercise.■

APPLICATIONS TO INTEGRAL FORMULAS IN VECTOR ANALYSIS. The preceding comparison between exterior differentiation and the ∇ operator leads to the following unified statement which includes the classical theorems of Green, Stokes and Gauss (also called the Divergence Theorem):

$$\int_{\mathrm{Bdy}(X)} \omega \quad = \quad \int_X d\omega$$

Here X is a region in $\mathbb{R}^2$ or $\mathbb{R}^3$ or an oriented surface, and $\mathrm{Bdy}(X)$ denotes its boundary curve(s) or surface(s).■

Proving this version of the theorems is beyond the scope of the course, but we have mentioned it to suggest the potential usefulness of differential forms for expressing somewhat complicated relationships in a relatively simple manner.

Connectedness

In many situations it is useful or necessary to assume that an open set has an additional property called connectedness.

Definition. Let $n = 2$ or 3 (actually, everything works for all $n \geq 2$, but in this course we are mainly interested in objects that exist in 2- or 3-dimensional space). An open subset U of $\mathbb{R}^n$ will be called a *connected open domain* if for each pair of points $\mathbf{p}$ and $\mathbf{q}$ in U there is a piecwise smooth curve Γ defined on $[0, 1]$ and taking values entirely in U such that $\Gamma(0) = \mathbf{p}$ and $\Gamma(1) = \mathbf{q}$.

Most examples of open sets in this course are either connected or split naturally into a finite union of pairwise disjoint open subsets. Here are some examples:

Example 1. An open disk of radius $r > 0$ about a point $\mathbf{p}$, consisting of all $\mathbf{x}$ such that $|\mathbf{x} - \mathbf{p}| < r$ is connected. If $\mathbf{x}$ and $\mathbf{y}$ belong to such a disk, then consider the *line segment* curve $\gamma(t) = t\,\mathbf{y} + (1-t)\,\mathbf{x}$, where $t \in [0,1]$. This is an infinitely differentiable curve (its coordinate functions are first degree polynomials), it joints $\mathbf{x}$ to $\mathbf{y}$, and we have

$$|\gamma(t)| \quad \leq \quad t\,|\mathbf{x}| \;+\; (1-t)\,|\mathbf{y}| \quad < \quad t\,r \;+\; (1-t)\,r \quad = \quad r$$

so that $\gamma(t)$ lies in the open disk of radius r for all $t \in [0,1]$.

Example 2. Let i be a number between 1 and n, and let H_i be the set of all points in $\mathbb{R}^n$ whose I^{th} coordinate satisfies $x_i \neq 0$. Then H_i splits into a union of the two sets H_i^+ and H_i^- of points where x_i is positive and negative respectively. Each of these is connected, and in fact two points in H_i^+ or H_i^- can be joined by the same sort of line segment curve as in Example 1. The reason for this is that if the i^{th} coordinates of $\mathbf{x}$ and $\mathbf{y}$ are positive or negative, the corresponding property holds for each point $\gamma(t)$.

Note, however, that H itself is **not** a connected open domain. Specifically, there is no curve joining the unit vector $\mathbf{e}_i$ to its negative. If such a curve did exist, then its i^{th} coordinate z_i would be a continuous function from $[0,1]$ to the reals such that $z_i(0) = -1$ and $z_i(1) = 1$. By the Intermediate Value Property for continuous functions on an interval, there would have to be some parameter value u for which $z_i(u) = 0$; but this would mean that $\gamma(u)$ could not belong to H_i, so we have a contradiction. The problem arises from our assumption that there was a continuous curve in H_i joining the two vectors in question, so no such curve can exist.■

To illustrate the role of connectedness, we shall consider the following question: *Suppose that U is an open subset of $\mathbb{R}^2$ or $\mathbb{R}^3$ and f is a real valued function on U such that all the partial derivatives of f are defined and equal to zero. Is f a constant function?*

The answer to this question depends upon whether or not U is connected.

Example. Let $U = H_i$, and defined f such that $f(\mathbf{x}) = 1$ if the coordinate x_i is positive and $f(\mathbf{x}) = -1$ if the coordinate x_i is negative. Then f is not constant but one can check directly that the partial derivatives of f are always defined and zero.

THEOREM. *Let U be a connected subset of $\mathbb{R}^2$ or $\mathbb{R}^3$, and let f be a real valued function on U such that all the partial derivatives of f are defined and equal to zero. Then f is a constant function.*

Proof. The key step is to prove the following: *Suppose that $\mathbf{p}$ and $\mathbf{q}$ are points in U such that the line segment joining $\mathbf{p}$ to $\mathbf{q}$ lies entirely in U. Then $f(\mathbf{p}) = f(\mathbf{q})$.*

To prove this, let $\mathbf{v} = \mathbf{q} - \mathbf{p}$, so that the line segment joining $\mathbf{p}$ to $\mathbf{q}$ has the parametrization $\gamma(t) = \mathbf{p} + t\,\mathbf{v}$. For each index i let v_i denote the i^{th} coordinate of $\mathbf{v}$. Let $g(t) = f\big(\gamma(t)\big)$; by the Chain Rule we have

$$g'(t) \quad = \quad \sum_i \frac{\partial f}{\partial x_i}\,\big(\,\gamma(t)\,\big) \cdot v_i$$

and the right hand side is zero because all the partial derivatives of f are zero. Since $g' = 0$, by results from single variable calculus we know that g is constant, and this means that $f(\mathbf{p}) = g(0) = g(1) = f(\mathbf{q})$.

To prove the theorem, suppose that $\mathbf{p}$ and $\mathbf{q}$ are arbitrary points in U. By the definition of connectedness there is a broken line curve joining these points. Suppose that this broken line curve consists of the line segments $S_1, \;\cdots\;, S_m$ such that $\mathbf{p} = \mathbf{x}_0$, $\mathbf{q} = \mathbf{x}_m$, and the endpoints for each S_j are given by $\mathbf{x}_{i-1}$ and $\mathbf{x}_i$. Then by the reasoning of the previous paragraph we know that $f(\mathbf{p}) = f(\mathbf{x}_0) = f(\mathbf{x}_1)$, $f(\mathbf{x}_1) = f(\mathbf{x}_2)$, and similarly all the values $f(\mathbf{x}_i)$ are all equal to each other. In particular, it follows that $f(\mathbf{p}) = f(\mathbf{x}_0) = \;...\; = f(\mathbf{x}_m) = f(\mathbf{q})$.■

II.2 : Smooth mappings

(Lipschutz, Chapters 6–7)

From a purely formal viewpoint, the generalization from real valued functions of several variables to vector valued functions is simple. An n-dimensional vector valued function is specified by its n coordinates, each of which is a real valued function. As in the case of one variable functions, a vector valued function is continuous if and only if each coordinate function is continuous.

One reason for interest in vector valued functions of several real variables is their interpretation as *geometric transformations*, which map geometric figures in the domain of definition to geometric figures in the target space of the function. For example, in linear algebra one has *linear transformations* given by homogeneous linear polynomials in the coordinates, and it is often interesting or useful to understand how familiar geometric figures in $\mathbb{R}^2$ or $\mathbb{R}^3$ are moved, bent or otherwise distorted by a linear transformation. Examples are discussed in most linear algebra texts (for example, see Section 2.4 of Fraleigh and Beauregard, *Linear Algebra*), and the following interactive wev site allows the user to view the images of various quadrilaterals under linear transformations, where the user has a wide range of choices for both geometric figure and the transformation:

`http://merganser.math.gvsu.edu/david/linear/linear.html`

The notion of a geometric mapping is also central to change of variables problems in multivariable calculus. For example, it one wants to evaluate a double integral over a region A in the Cartesian coordinate plane using polar coordinates, it is necessary to understand the geometric figure B in the plane that maps to A under the vector valued function of two variables

$$\mathbf{Cart}(r,\,\theta) \quad = \quad (r\,\cos\theta,\, r\,\sin\theta)\ .$$

Since many different sets of polar coordinates yield the same point in Cartesian coordinates, it is generally appropriate to assume that B lies in some set for which Cartesian coordinates are unique or almost always so. For example, one might take B to be the set of all points that map to A and whose r and θ coordinates satisfy $0 \leq r$ and $0 \leq \theta \leq 2\pi$. Some illustrations appear in the following site; the collection of pictures in the first is particularly extensive and makes very effective use of different colors.

`http://loriweb.pair.com/8polarcoord1.html`

`omega.albany.edu:8008/calc3/double-integrals-dir/polar-coord-m2h.html`

If a vector valued function of several variables is defined on a connected domain in some $\mathbb{R}^n$, then one can formulate a notion of partial derivatives using the coordinate functions and the usual methods of multivariable calculus, but exactly as in that subject such partial derivatives can behave somewhat erratically if they are not continuous. However, if these partial derivatives are continuous, then one has the following critically important generalization of a basic result on real valued functions of several variables:

LINEAR APPROXIMATION PROPERTY. *Suppose that U is a connected domain in $\mathbb{R}^n$ and that $f : U \to \mathbb{R}^m$ is a function with continuous first partial derivatives on U. Denote the coordinate functions of f by f_i, and for each $\mathbf{x} \in U$ let $Df(x)$ be the matrix whose i^{th} row is given by the gradient vector $\nabla f_i(\mathbf{x})$. Then for all sufficiently small but nonzero vectors $\mathbf{h} \in \mathbb{R}^n$ we have*

$$f(\mathbf{x}+\mathbf{h}) \quad = \quad f(\mathbf{x}) \ + \ \big[Df(\mathbf{x})\big]\mathbf{h} \ + \ |\mathbf{h}|\,\theta(\mathbf{h})$$

where $\theta(\mathbf{h})$ satisfies

$$\lim_{\mathbf{h}\to\mathbf{0}} \theta(\mathbf{h}) \quad = \quad \mathbf{0} \, .$$

The matrix $Df(\mathbf{x})$ is often called the *derivative* of f at $\mathbf{x}$. If $n = m$ then the determinant of this matrix is just the *Jacobian* which arises in the change of variables formula for multiple integrals.

Sketch of proof. For scalar valued functions, a version of this result is established in multivariable calculus; specifically, in our case this result says that the coordinate functions satisfy equations of the form

$$f_i(\mathbf{x}+\mathbf{h}) \quad = \quad f_i(\mathbf{x}) \; + \; \nabla f_i(\mathbf{x}) \cdot \mathbf{h} \; + \; |\mathbf{h}| \, \theta(_i\mathbf{h})$$

where $\theta(\mathbf{h})$ satisfies

$$\lim_{\mathbf{h}\to\mathbf{0}} \theta_i(\mathbf{h}) \quad = \quad \mathbf{0} \, .$$

By construction, the rows of $Df(\mathbf{x})$ are the gradient vectors of the coordinate functions at $\mathbf{x}$, and consequently the coordinates of $\big[Df(\mathbf{x})\big]\mathbf{h}$ are given by the expressions $\nabla f_i(\mathbf{x}) \cdot \mathbf{h}$. The function $\theta(\mathbf{h})$ is defined so that it coordinates are the functions $\theta_i(\mathbf{h})$, and the limit of θ at $\mathbf{0}$ is $\mathbf{0}$ because the limit of each θ_i at $\mathbf{0}$ is 0.■

The preceding result implies that a vector valued function of several variables with continuous partial derivatives has a well behaved first degree approximation by a function of the form

$$g(\mathbf{x}+\mathbf{h}) \quad = \quad g(\mathbf{x}) \; + \; B\,\mathbf{h}$$

for some $m \times n$ matrix B (namely, the derivative matrix).

WARNING. Frequently mathematicians and physicists use *superscripts* to denote coordinates. Of course this conflicts with the usual usage of superscripts for exponents, so one must be aware that superscripts may be used as indexing variables sometimes. Normally such usage can be detected by the large number of superscripts that appear or their use in places where one would normally not expect to see exponents.

Smoothness classes. As for functions of one variable, we say that a vector valued function of several variables is smooth of class $\mathcal{C}^r$ if its coordinate functions have continuous partial derivatives of order $\leq r$ (agreeing that $\mathcal{C}^0$ means continuous) and that a function is smooth of class $\mathcal{C}^\infty$ if its coordinate functions have continuous partial derivatives of all orders.

The concept of derivative matrix for a vector valued function leads to a very neat formulation of the Chain Rule:

VECTOR MULTIVARIABLE CHAIN RULE. *Let U and V be connected domains in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, let $f : U \to V$ be a map whose coordinate functions have continuous partial derivatives at $\mathbf{x}$, and let $g : V \to \mathbb{R}^p$ be a map whose coordinate functions have continuous partial derivatives at $f(\mathbf{x})$. Then the composite $g \circ f$ defined by*

$$g \circ f(\mathbf{y}) \quad = \quad g\big(\, f(\mathbf{y})\,\big)$$

also has coordinates with continuous partial derivatives at $\mathbf{x}$ and

$$D\,[g \circ f]\,(\mathbf{x}) = D(g)\,(f(\mathbf{x}))\, \circ Df(\mathbf{x}) \, .$$

Proof. This follows directly by applying the chain rule for scalar valued functions to the partial derivatives of the coordinate functions for $g \circ f$.■

COROLLARY. *In the preceding result, if f and g are smooth of class $\mathcal{C}^r$, then the same condition holds for their composite $g \circ f$.*

Proof. First of all, if the result can be shown for $r < \infty$ the case $r = \infty$ will follow out because $\mathcal{C}^\infty$ is equivalent to $\mathcal{C}^s$ for all $s < \infty$. Therefore we shall assume $r < \infty$ for the rest of the proof.

If h is a q-dimensional vector valued function of p variables of class $\mathcal{C}^r$, then the derivative matrix of h may be viewed as a $p \times q$ matrix valued function of p variables, or equivalently as a pq-dimensional vector valued function of p variables, and this function is smooth of class $\mathcal{C}^{r-1}$. We shall use this fact to prove the corollary by induction on r.

Suppose first that $r = 1$. Then the Chain Rule states that the entries of $D\,[g \circ f]\,(\mathbf{x})$ are polynomials in the entries of $D(g)\,(f(\mathbf{x}))$ and $Df(\mathbf{x})$. Since Dg, Df and f are all continuous and a composite of continuous functions is continuous, it follows that $D\,[g \circ f]\,(\mathbf{x})$ is a continuous function of $\mathbf{x}$.

Suppose now that we know the result for $s < r$, where $r \geq 2$. Then exactly the same sort of argument applies, with $\mathcal{C}^{r-1}$ replacing "continuous" in the final sentence; this step is justified by the induction hypothesis.■

The file `changevarexamples.pdf` (as usual in the course directory) describes some examples of smooth transformations f from $\mathbb{R}^2$ to itself.

II.3 : Inverse and Implicit Function Theorems

(Lipschutz, Chapter 7)

The following topics are often discussed very rapidly or not at all in multivariable calculus courses, but we shall need them at many points in the discussion of surfaces. The text for the Department's courses on single and multivariable calculus courses (Colley, *Multivariable Calculus*) discusses these results as an optional part of Section 2.6 on pages 162–167. More detailed statements and proofs of the results are contained in the text for the Department's advanced undergraduate course on real variables (Rudin, *Principles of Mathematical Analysis*, Third Edition). A statement of the one result (the Inverse Function Theorem) also appears on page 131 of DO CARMO. Here are some online references:

`http://www.ualberta.ca/MATH/gauss/fcm/calculus/` (*continue with next line*)
`multvrbl/basic/ImplctFnctns/invrs_fnctn_explntn_illstrtn2.gif`

`http://artsci.wustl.edu/~e4111jn/InvFT14.pdf`

`http://www.sas.upenn.edu/~kim37/mathcamp/Eduardo_inverse.pdf`

`http://en.wikipedia.org/wiki/Inverse_function_theorem`

We shall begin our discussion with the Implicit Function Theorem. The simplest form of this result is generally discussed in the courses on differential calculus. In these courses one assumes that some equation of the form $F(x,y)=0$ can be solved for y as a function of x and then attempts to find the derivative y'. The standard formula for the latter is

$$\frac{df}{dx} \quad = \quad -\frac{\left(\frac{\partial F}{\partial x}\right)}{\left(\frac{\partial F}{\partial y}\right)}$$

where of course this formula can be used only if the denominator is nonzero. In fact if we have a point (a,b) such that $F(a,b)=0$ and the second partial of F at (a,b) is not zero, then the simplest case of the Implicit Function Theorem proves that one can indeed find a differentiable function $f(x)$ for all values of x sufficiently close to a such that $f(a)=b$ and for all nearby values of x we have

$$y=f(x) \iff F(x,\,y)=0\ .$$

Here is a general version of this result:

IMPLICIT FUNCTION THEOREM. *Let U and V be connected domains in $\mathbb{R}^n$ and $\mathbb{R}^m$ respecitvely, and let $f:U\times V\to\mathbb{R}^m$ be a smooth function such that for some $\mathbf{p}=(\mathbf{a},\,\mathbf{b})\in U\times V$ we have $f(\mathbf{a},\,\mathbf{b})=0$ and the partial derivative of f with respect to the last m coordinates is invertible. Then there is an $r>0$ and a smooth function*

$$g:N_r(\mathbf{p})\to V$$

such that $g(\mathbf{a})=\mathbf{b}$ and for all $u\in U_0$ we have $f(\mathbf{u},\mathbf{v})=0$ if and only if $\mathbf{v}=g(\mathbf{u})$.■

EXPLANATIONS. **(1)** We view the Cartesian product $U\times V$ as a subset of $\mathbb{R}^{n+m}$ under the standard identification of the latter with $\mathbb{R}^n\times\mathbb{R}^m$.

(2) The partial derivative of f with respect to the last m coordinates is the derivative of the function $f^*(v)=f(x,v)$, and smooth means smooth of class $\mathcal{C}^r$ for some r such that $1\leq r\leq\infty$.

Although it is possible to prove simple cases of this result fairly directly, the usual way of establishing the Implicit Function Theorem is to derive it as a consequence of another important result known as the *Inverse Function Theorem.* We shall be using this result extensively throughout the remainder of the course.

Once again it is instructive to recall the special case of this result that appears in single variable calculus courses. For real valued functions on an interval, the Intermediate Value Property from elementary calculus implies that local inverses exist for functions that are strictly increasing or strictly decreasing. Since the latter happens if the function has a derivative that is everywhere positive or negative close to a given point, one can use the derivative to recognize very quickly whether local inverses exist in many cases, and in these cases one can even compute the derivative of the inverse function using the standard formula:

$$g \quad = \quad f^{-1} \qquad \Longrightarrow \qquad g'(y) \quad = \quad \frac{1}{f'(g(y))}$$

Of course this formula requires that the derivative of f is not zero at the points under consideration.

If we are dealing with a function of n variables whose values are given by n-dimensional vectors, one has the following far-reaching generalization in which the nonvanishing of the derivative is replaced by the invertibility of the derivative matrix, or equivalently by the nonvanishing of the Jacobian:

INVERSE FUNCTION THEOREM. *Let U be a connected domain in $\mathbb{R}^n$, let $a \in U$, and let $f : U \to \mathbb{R}^n$ be a $\mathcal{C}^r$ map (where $1 \le r \le \infty$) such that $Df(\mathbf{a})$ is invertible. Then there is a connected domain $W \subset U$ containing* $\mathbf{a}$ *such that the following hold:*

(i) The restriction of f to W is $1-1$ and its image is a connected domain V.

(ii) There is a $\mathcal{C}^r$ inverse map g from V to some connected domain $U_0 \subset U$ containing $\mathbf{a}$ *such that $g(f(\mathbf{x})) = \mathbf{x}$ on U_0.*■

For the purposes of this course it will suffice to understand the statements of the Inverse and Implicit Function Theorems, so we shall restrict attention to this point and refer the reader to Rudin for detailed proofs; a similar treatment of this material appears in Section II.2 of the following set of notes for another course that are available online:

`http://www.math.ucr.edu/~res/math205C/lectnotes.pdf`

.

WARNING. The Inverse Function only implies the existence of an inverse and does not provide any general method for expressing the inverse in terms of the functions studied in first year calculus, even if the coordinates of the original function have such a form. In fact, this is true even if we restrict attention to real valued functions of a single real variable. One example of such a function is the inverse function to $f(x) = x + e^x$; note that this function is strictly increasing since its derivative is always positive, and its limits as $x \to \pm\infty$ are equal to $\pm\infty$ respectively. Therefore it follows that f has a strictly increasing inverse function, but it turns out that this function cannot be expressed in terms of the functions one encounters in first year calculus. The online documents

`http://math.ucr.edu/~res/math205A/Lambertfcn.pdf`

`http://apmathw.uwo.ca/~djeffrey/offprints/ITSF2006.pdf`

provide some information and further references for this example and the closely related *Lambert W-function*, which is defined by the identity

$$z \quad = \quad W(z)\, e^{W(z)} \ .$$

Another noteworthy example of a simply described function with a relatively nonsimple inverse is given by $f(x) = x^5 + x^3 + x$. This function is also strictly increasing, and its limits as $x \to \pm\infty$ are equal to $\pm\infty$ respectively, so that an inverse function exists. However, this function cannot be expressed in familiar sorts of terms using addition, subtraction, multiplication, division, and taking n^{th} roots for $n \le 5$. Further discussion of this example appears in Section II.3 of the following document:

`http://math.ucr.edu/~res/math144/transcendentals.pdf`

On a more positive note, the functions $x+e^x$ and x^5+x^3+x have convergent power series expansions for all real values of x (of course, in the second case there are only finitely many nonzero terms), and for each example and each real number a the inverse function also has convergent power series expansions at $x = a$; these formulas are valid over suitable open intervals centered at a of the form $(a - R_a, a + R_a)$ for some $R_a > 0$, but these expansions are not valid over the entire real line. In fact, one can set up equations for the power series coefficients of the inverse functions in terms of the coefficients for the power series of original functions at $x = a$ (see the file `inverse-series.pdf` in the course directory for more information).

REMARKS ON PROOFS. Finally, here are online references for the proofs of the Inverse and Implicit Function Theorems. These are similar to the proofs in the previous online reference for the theorems studied in the present section.

http://planetmath.org/encyclopedia/ProofOfInverseFunctionTheorem.html

http://planetmath.org/encyclopedia/ProofOfImplicitFunctionTheorem.html

Change of variables in multiple integrals

In multivariable calculus courses, one is interested in changes of variables arising from smooth mappings that are 1–1 and onto with Jacobians that are nonzero "almost everywhere." The standard polar, cylindrical and spherical coordinates are the most basic examples provided that one restricts the angle parameters θ and ϕ (in the spherical case) so there is no ambiguity; the Jacobian condition is reflected by the fact that this quantity is nonzero for polar and cylindrical coordinates if $r \neq 0$, and it is nonzero for spherical coordinates so long as $\rho^2 \sin\phi \neq 0$. Further discussion of this result in the general case appears on pages 333–336 of the background reference text by Marsden, Tromba and Weinstein, and on pages 995–1001 of the background reference text by Larson, Hostetler and Edwards. Exercises 37–40 on page 339 of the first reference and exercises 60–61 on page 1004 of the second are recommended as review. Other possible sources for background include Section 5.5 of COLLEY and the following online commentary regarding the latter:

http://math.ucr.edu/∼res/math10B/comments0505.pdf

For the sake of completeness, here is a statement of the basic formula that applies to all dimensions (not just 2 and 3).

CHANGE OF VARIABLES FORMULA. *Let U and V be connected domains in $\mathbb{R}^n$, and let $f : U \to V$ be a map with continuous partial derivatives that is $1-1$ onto has a nonzero Jacobian everywhere. Suppose that A and B are "nice" subsets of U and V respectively that correspond under f, and let h be a continuous real valued function on V. Then we have*

$$\int_B h(\mathbf{v})\,d\mathbf{v} \quad = \quad \int_A h\big(\,f(\mathbf{u})\,\big)\,|\det Df(\mathbf{u})|\,d\mathbf{u}\ .\blacksquare$$

As in the case of polar, cylindrical and spherical coordinates, the result still holds if the Jacobian vanishes on a set of points that is not significant for computing integrals (in the previous terminology, one needs that the Jacobian is nonzero "almost everywhere," and this will happen if the zero set of the Jacobian is defined by reasonable sets of equations).

One can weaken the continuity assumption on h even more drastically, but this requires a more detailed insights into integrals than we need here.■

There is an extensive discussion of the proof of this result along with some illustrative examples in Section IV.5 of the book *Advanced Calculus of Several Variables*, by C. H. Edwards, and a mathematically complete proof appears on pages 252–253 of the previously cited book by Rudin. As noted on page 252 of Rudin, this form of the change of variables theorem is too restrictive for some applications, but in most of the usual applications one can modify the proof so that it extends to somewhat more general situations; generally the necessary changes are relatively straightforward, but carrying out all the details can be a lengthy process.

Remark on the absolute value signs. In view of the usual change of variables formulas for ordinary integrals in single variable calculus, it might seem surprising that one must take the absolute value of the Jacobian rather than the Jacobian itself. Some comments about the reasons for this are given in the middle of page 252 in Rudin's book. In fact, we dealt specifically with this issue in Section I.3, when we proved that arc length remains unchanged under reparametrization.

II.4 : Congruence of geometric objects

(Lipschutz, Chapter 6)

The notion of congruence for geometrical figures plays a central role in classical synthetic Euclidean geometry. For some time mathematicians — and users of mathematics — have generally studied geometrical questions analytically using vectors and linear algebra (these often provide neat and efficient ways of managing the usual coordinates in analytic geometry). A few simple examples often appear in introductory treatments of vectors in calculus books or elsewhere, and in fact one can state and prove everything in classical Euclidean geometry by such analytic means. However, there are still numerous instances where it is useful to employ ideas from classical synthetic geometry, and in particular this is true in connection with the Frenet-Serret Formulas from Unit I. Therefore we shall formulate the analytic notion of congruence rigorously, and we shall use it to state an important congruence principle for differentiable curves.

Isometries of $\mathbb{R}^n$

Definition. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a mapping (with no assumptions about continuity or differentiability). Then f is said to be an **isometry** of $\mathbb{R}^n$ if it is a 1–1 correspondence from $\mathbb{R}^n$ onto itself such that

$$|f(\mathbf{x}) \; - \; f(\mathbf{y})| \quad = \quad |\mathbf{x} \; - \; \mathbf{y}|$$

for all $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$.

Two subsets $A, B \subset \mathbb{R}^n$ are said to be **weakly congruent** if there is an isometry f of $\mathbb{R}^n$ such that B is the image of A under the mapping f. If A and B are weakly congruent, then one often writes $A \cong B$ in the classical tradition.

Since inverses and composites of isometries are isometries (and the identity is an isometry), it follows that weak congruence is an equivalence relation.

The first step is to prove the characterization of isometries of a finite-dimensional Euclidean space that is often given in linear algebra textbooks. To simplify our notation, we shall use the term *finite-dimensional Euclidean space* to denote the vector spaces $\mathbb{R}^n$ with their standard inner products.

PROPOSITION. *If* $\mathbf{E}$ *is a finite-dimensional Euclidean space and* F *is an isometry from* $\mathbf{E}$ *to itself, then* F *may be expressed in the form* $F(\mathbf{x}) = \mathbf{b} + A(\mathbf{x})$ *where* $\mathbf{b} \in E$ *is some fixed vector and* A *is an orthogonal linear tranformation of* $\mathbf{E}$ *(i.e., in matrix form we have that* ${}^{\mathbf{T}}A = A^{-1}$ *where* ${}^{\mathbf{T}}A$ *denotes the transpose of* A*).*

Notes. It is an elementary exercise to verify that the composite of two isometries is an isometry (and the inverse of an isometry is an isometry). If A is orthogonal, then it is elementary to prove that $F(\mathbf{x}) = \mathbf{b} + A(\mathbf{x})$ is an isometry, and in fact this is done in most if not all undergraduate linear algebra texts. On the other hand, if $A = I$ then the map above reduces to a **translation** of the form $F(\mathbf{x}) = \mathbf{b} + \mathbf{x}$, and such maps are isometries because they satisfy the even stronger identity

$$F\big(\mathbf{x} - \mathbf{y}\big) \quad = \quad \mathbf{x} \; - \; \mathbf{y} \; .$$

Therefore every map of the form $F(\mathbf{x}) = \mathbf{b} + A(\mathbf{x})$, where $\mathbf{b} \in E$ is some fixed vector and A is an orthogonal linear tranformation of $\mathbf{E}$, is an isometry of $\mathbf{E}$. Therefore the proposition gives a complete characterization of all isometries of $\mathbf{E}$.

Sketch of proof. This argument is often given in linear algebra texts, and if this is not done then hints are frequently given in the exercises, so we shall merely indicate the basic steps.

First of all, the set of all isometries of $\mathbf{E}$ is a group (sometimes called the *Galileo group* of $\mathbf{E}$). It contains both the subgroups of orthogonal matrices and the subgroup of translations ($G(\mathbf{x}) = \mathbf{x}+\mathbf{c}$ for some fixed vector $\mathbf{c}$), which is isomorphic as an additive group to $\mathbf{E}$ with the vector addition operation. Given $b \in \mathbf{E}$ let $\mathbf{S_b}$ be translation by $\mathbf{b}$, so that $A = \mathbf{S}_{-F(\mathbf{0})} \circ F$ is an isometry from $\mathbf{E}$ to itself satisfying $G(\mathbf{0}) = \mathbf{0}$. If we can show that G is linear, then it will follow that G is given by an orthogonal matrix and the proof will be complete.

Since G is an isometry it follows that

$$\left|G(\mathbf{x}) - G(\mathbf{y})\right|^2 \quad = \quad \left|\mathbf{x} - \mathbf{y}\right|^2$$

and since $G(0) = 0$ it also follows that g is length preserving. If we combine these special cases with the general formula displayed above we conclude that $\langle G(\mathbf{x}), G(\mathbf{y})\rangle = \langle \mathbf{x}, \mathbf{y}\rangle$ for all $\mathbf{x},\ \mathbf{y} \in \mathbf{E}$. In particular, it follows that G sends orthonormal bases to orthonormal bases. Let $\{\mathbf{u}_1,\ \cdots\ , \mathbf{u}_n\}$ be an orthonormal basis; then we have

$$\mathbf{x} \quad = \quad \sum_{i=1}^{n} \langle \mathbf{x}, \mathbf{u}_i\rangle \cdot \mathbf{u}_i$$

and likewise we have

$$G(\mathbf{x}) \quad = \quad \sum_{i=1}^{n} \langle G(\mathbf{x}), G(\mathbf{u}_i)\rangle \cdot G(\mathbf{u}_i) \ .$$

Since G preserves inner products we know that

$$\langle \mathbf{x}, \mathbf{u}_i\rangle \quad = \quad \langle G(\mathbf{x}), G(\mathbf{u}_i)\rangle \cdot G(\mathbf{u}_i)$$

for all i, and this implies that G is a linear transformation.■

Since an isometry is a mapping from $\mathbb{R}^n$ to itself, it is meaningful to ask about its continuity or differentiability properties. The following result answers such questions simply and completely.

PROPOSITION. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a mapping of the form $F(\mathbf{x}) = \mathbf{b} + A(\mathbf{x})$, where $\mathbf{b} \in \mathbb{R}^n$ is some fixed vector and A is an arbitrary square matrix. Then for all $\mathbf{x} \in \mathbb{R}^n$ we have $DF(\mathbf{x}) = A$.*

COROLLARY. *Let V be open in $\mathbb{R}^m$, let $g : V \to \mathbb{R}^m$ have a continuous derivative, and let A be an $n \times n$ matrix; by an abuse of language, let A also denote the linear transformation from $\mathbb{R}^n$ to itself defined via left multiplication by A. Then we have $D(A \circ g) = A \circ Dg$.*

Proofs. The statement in the proposition follows from the definition of the derivative as a matrix whose entries are the partial derivatives of the coordinate functions. In this case the coordinate functions are all first degree polynomials in n variables. The statement in the corollary follows from the proposition and the Chain Rule.■

The concept of weak congruence is close, but not identical, to the idea that there is a dynamic rigid motion taking one figure to another; the main difference is that weak congruence also allows the possibility that one figure is a mirror image of the other. For our purposes it is enough to know that if F is an isometry then the orthogonal linear transformation DF has determinant equal to ± 1, and the intuitive concept of rigid motion corresponds to the case where the determinant is equal to $+1$. Therefore we shall say that F is a *rigid motion* if this determinant is $+1$, and we shall say that two weakly congruent figures A and B are *strongly congruent*, or more simply **congruent**, if there is a rigid motion taking one to the other.

The file `congruence000.pdf` describes the relation of congruence or isometry in these notes to congruence in elementary geometry more explicitly; also, the file

`http://math.ucr.edu/~res/math133/metgeom.pdf`

discusses the relationship of linear algebra to elementary geometrical congruence and similarity in more detail (but at a somewhat higher level).

Congruence and differentiable curves

We shall say that two continuous curves $\alpha, \beta : [a, b] \to \mathbb{R}^n$ are **congruent** if there is an isometry F of $\mathbb{R}^n$ such that $\beta = F \circ \alpha$. We are interested in the relationship between the curvatures and torsions of congruent curves.

PROPOSITION. *Let $\alpha, \beta : [a, b] \to \mathbb{R}^3$ be congruent differentiable curves whose tangent vectors have constant length equal to 1 and whose curvatures are never zero. Then the curvature and torsion functions for α and β are equal.*

Proof. Let F be a rigid motion of $\mathbb{R}^3$ such that $\beta = F \circ \alpha$, express F in the usual form $F(\mathbf{x}) = \mathbf{b} + A(\mathbf{x})$ where $\mathbf{b} \in \mathbb{R}^3$ and A is an orthogonal transformation whose determinant is equal to $+1$, and suppose that α has k continuous derivatives. By the Chain Rule we know that β also has k continuous derivatives, and in fact $\beta^{(k)} = A \circ \alpha^{(k)}$.

Since $|\beta'| = |\alpha'| = 1$, it follows that the curvatures are given by $\kappa_\alpha = |\alpha''|$ and $\kappa_\beta = |\beta''|$. Since $\beta'' = A \circ \alpha''$ and A is orthogonal, it follows that $|\beta''| = |\alpha''|$, and hence the curvatures of α and β are equal.

We shall now show that the Frenet trihedra for the curves are related by

$$\Big(\mathbf{T}_\beta, \mathbf{N}_\beta, \mathbf{B}_\beta\Big) \quad = \quad \Big(A(\mathbf{T}_\alpha), A(\mathbf{N}_\alpha), A(\mathbf{B}_\alpha)\Big) \ .$$

The result for the unit tangent vector is just a restatement of the relationship $\beta' = A \circ \alpha'$, and the result for the principal unit normal follows because we have

$$\mathbf{N}_\beta \quad = \quad \frac{1}{|\beta''|}\,\beta'' \quad = \quad \frac{1}{|\beta''|}\,A(\alpha'') \quad = \quad \frac{1}{|\alpha''|}\,A(\alpha'') \quad =$$

$$A\left(\frac{1}{|\alpha''|}\,\alpha''\right) \quad = \quad A\big(\mathbf{N}_\alpha\big) \ .$$

We must next compare the binormals; this amounts to checking whether the following cross product formula holds:

$$A\big(\mathbf{B}_\alpha\big) \quad = \quad A\big(\mathbf{T}_\alpha\big) \ \times \ A\big(\mathbf{N}_\alpha\big) \quad = \quad \mathbf{T}_\alpha \ \times \ \mathbf{N}_\beta$$

We shall do this using the Recognition Formula from Section I.1. By that result, all we have to check is that the triple product satisfies

$$\big[A(\mathbf{T}_\alpha), A(\mathbf{N}_\alpha), A(\mathbf{B}_\alpha)\big] \quad = \quad +1 \ .$$

This triple product is just the determinant of the matrix whose columns are the three vectors. This matrix in turn factors as a product of A and the matrix whose columns are the Frenet trihedron for α, and by the multiplicative properties of determinants we then have

$$\big[A(\mathbf{T}_\alpha),\, A(\mathbf{N}_\alpha),\, A(\mathbf{B}_\alpha)\big] \quad = \quad \det A \cdot \big[\mathbf{T}_\alpha,\, \mathbf{N}_\alpha,\, \mathbf{B}_\alpha\big] \quad = \quad (+1)\cdot(+1) \quad = \quad +1$$

so that the Recognition Formula implies the cross product identity. This completes the verification of the relationship between the Frenet trihedra.

To complete the proof we need to show that the torsions satisfy $\tau_\beta = \tau_\alpha$. By definition we have $\tau_\beta(s) = -\mathbf{B}_\beta{}'(s)\cdot\mathbf{N}_\beta(s)$. Since $\mathbf{B}_\beta = A(\mathbf{B}_\alpha)$, there is a corresponding identity involving derivatives, and therefore by the preceding paragraph we have

$$\tau_\beta(s) \quad = \quad -A\big(\mathbf{B}_\alpha{}'(s)\big)\cdot A\big(\mathbf{N}_\alpha(s)\big) \ .$$

Since A is orthogonal, it preserves inner products, and consequently the right hand side is equal to $-\mathbf{B}_\alpha{}'(s)\cdot\mathbf{N}_\alpha(s)$, which by definition is just $\tau_\alpha(s)$. Combining these observations, we see that the torsions of α and β are equal as claimed.■

Uniqueness up to congruence

We are now ready to prove that curvature and torsion often determine a differentiable curve up to congruence.

UNIQUENESS UP TO CONGRUENCE. *Let α and β be sufficiently differentiable curves in $\mathbb{R}^3$ defined on the same open interval J containing s_0, and assume that their curvatures and torsions satisfy $\kappa_\alpha = \kappa_\beta > 0$ and $\tau_\alpha = \tau_\beta$. Then there is an isometry F of $\mathbb{R}^3$ such that $\det DF(\mathbf{x}) = +1$ for all $\mathbf{x}$ and $\beta = F\circ\alpha$.*

Proof. Let $\big(\mathbf{T}_x,\, \mathbf{N}_x,\, \mathbf{B}_x\big)$ be the Frenet trihedron for the curve $x = \alpha$ or β at parameter value s_0. If P and Q denote the matrices whose columns are given by $\big\{\mathbf{T}_\alpha,\, \mathbf{N}_\alpha,\, \mathbf{B}_\alpha\big\}$ and $\big\{\mathbf{T}_\beta,\, \mathbf{N}_\beta,\, \mathbf{B}_\beta\big\}$ respectively, then P and Q are orthogonal matrices with determinants equal to $+1$ (this follows because the columns are orthonormal and the third is the cross product of the first two). Therefore the matrix $C = P\,Q^{-1}$ is also orthogonal with determinant equal to $+1$. If we define F by the formula

$$f(\mathbf{x}) \quad = \quad C(\mathbf{x}) \ + \ (\beta(s_0) - \alpha(s_0)\,)$$

then $\gamma = F\circ\alpha$ is a curve whose curvatures and torsions are equal to those of α and β, and and its Frenet trihedron at parameter value s_0 is equal to the corresponding trihedron for β. By the local uniqueness portion of the Fundamental Theorem of Local Curve Theory, it follows that there is an open subinterval $J' \subset J$ containing s_0 such that the restrictions of $\gamma = F\circ\alpha$ and β to J' are equal.■

There is a more detailed proof of the preceding result in the file `curve-congruence.pdf`.

It is possible to prove a similar result on uniqueness up to congruence for plane curves with a given curvature function; as in the 2-dimensional versions of the result from Section I.5, there is no torsion function and it is not necessary to assume that the curvature is everywhere nonzero. The precise formulation of this result and its proof are left to the reader.■

III. Surfaces in 3-dimensional space

In Unit I we discussed two approaches to studying a curve, either by viewing it as a set of points in the plane or 3-dimensional space, or in terms of a parametrization. Similar considerations apply to surfaces in $\mathbb{R}^3$. Intuititvely speaking, a surface should be a subset that resembles a portion of the plane near every point, and this will be the case if we have a suitable description of the surface by parametric equations defined on some connected domain in $\mathbb{R}^2$. However, as noted on page 57 of DO CARMO, there is a major difference. For curves, it is often best simply to think of the curve in terms of the vector valued function given by a parametrization. On the other hand, for surfaces there is more of a balance between them as subsets of 3-dimensional space and objects given by their parametrizing functions. As noted on page *ix* of O'NEILL, a clear an adequate definition of surfaces is important, but this is not always given in the classical references; our definition will be equivalent to the ones in O'NEILL and DO CARMO.

One of the ultimate goals of classical surface theory is an analog of the Fundamental Theorem of Local Curve Theory, which states that many regular smooth curves in $\mathbb{R}^3$ are completely determined near a point by their curvatures and torsions. The corresponding result for surfaces may be viewed as a statement that a surface in $\mathbb{R}^3$ is determined by a pair of 2×2 matrix valued functions known as the *first and second fundamental forms*; in fact, both of these forms take values in the set of symmetric 2×2 matrices, and the possibilities for the first fundamental form are even more significantly restricted. This unit and the next one develop many of the basic concepts that are needed to study the differential geometry of surfaces, including some needed to formulate and to prove a fundamental theorem for local surface theory. As in the case of curves, much of the work involves generalizations of material from standard multivariable calculus courses. We shall not get to the fundamental theorem in this course, but there is a discussion of this result in Section V.2 of these notes.

III.1 : Mathematical descriptions of surfaces

(Lipschutz, Chapter 8)

> One weakness of classical differential geometry is its lack of any adequate definition of *surface*.
>
> O'NEILL, Preface, p. ix.

Some of the most basic examples of curves in $\mathbb{R}^2$ are given by the graphs of differentiable functions, and they can be described either as the set of points (x, y) where $y = f(x)$ or alternatively using a parametrization of the form $\mathbf{r}(t) = (t,\ f(t)\,)$. Likewise, some of the most basic examples of surfaces in $\mathbb{R}^3$ are given by the graphs of differentiable functions, and they can be described either as the set of points (x, y, z) where $z = f(x, y)$ or else by means of a parametrization $\mathbf{S}(u, v) = (u,\ v,\ f(u, v)\,)$.

If F is a function of two variables defined near (a, b) so that $F(a, b) = 0$ but the second partial derivative at (a, b) is nonzero, then the Implicit Function Theorem implies that locally one can solve the equation $F(x, y) = 0$ for y in terms of x, and it follows that locally the set $F(x, y) = 0$ is the image of a parametrized curve. More generally, if we know that $\nabla F(x, y) \neq \mathbf{0}$ whenever $F(x, y) = 0$, then at each point we can locally solve for one coordinate in terms of the other, and using these solutions one can generally find a parametrization of the level set defined by the equation $F(x, y) = 0$ which makes the latter into a regular smooth curve, at least if the level set consists of only one connected piece (this happens for the circle defined by $x^2 + y^2 = 1$ but not for the hyperbola $y^2 - x^2 = 1$). Proofs and more details about such constructions appear on pages 68–73 of THORPE (see pages 16 and 26 of the latter for some key definitions).

Similarly, if F is a function of three variables such that $\nabla F(x, y, z) \neq \mathbf{0}$ whenever $F(x, y) = 0$, then at each point we can locally solve for one coordinate in terms of the other two, so we have local parametrizations at each point. However, it is far more difficult to put together a global parametrization even if the level set defined by $F(x, y, z) = 0$ consists only of one connected piece. Perhaps the most basic example of this occurs for the unit sphere S^2, which corresponds to the equation $x^2 + y^2 + z^2 = 1$. It is easy to check the gradient condition for this example, and it is also easy to see write down explicit solutions for one variable in terms of the other two. However, it is not easy to write down a parametrization in elementary terms. The obvious parametrizations that one gets at different points cannot be pieced together as easily as one can piece together parametrizations for curves. In the case of curves, it is enough to match things up at boundary points of the intervals on which the partial parametrizations are defined , but the boundary sets for the two dimensional planar regions cannot be dealt with so easily. Another point to consider is that the parametrization of S^2 by spherical coordinates

$$\Sigma(\theta, \phi) \quad = \quad \big(\cos\theta \, \sin\phi, \, \sin\theta \, \sin\phi, \, \cos\phi\big)$$

is somewhat less regular than the corresponding parametrization of the unit circle as $(\cos\theta, \sin\theta)$ because it sends the infinite set of all parameter pairs with $\phi = 0$ to the north pole, and it also sends the infinite set of all parameter pairs with $\phi = \pi$ to the south pole. Just as we want parametrizations for curves that are regular in the sense that their derivatives are zero, we shall also want parametrizations for surfaces that are regular in the sense that every directional derivative at every point is nonzero.

These considerations suggest that we need more flexibility with surface parametrizations than we had for curve parametrizations. All of this will be made mathematically precise in the next section.

III.2 : Parametrizations of surfaces

(Lipschutz, Chapter 8)

The first objective is to define a regular smooth surface parametrization. This definition is very close to the definition of a regular smooth parametrization for a curve.

Definition. A *regular smooth surface parametrization* of class $r \geq 1$ is a smooth $\mathcal{C}^r$ map $\mathbf{x}$ from a connected domain U in $\mathbb{R}^2$ to $\mathbb{R}^3$ such that the 3×2 matrix $D\mathbf{x}(u,v)$ has maximum rank (which equals 2) for all $(u,v) \in U$.

The condition on the matrix is equivalent to the nonvanishing of the cross product of the partial derivative vectors

$$\frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v}$$

at all points of U, and in fact this is the form of the condition that is most often used in the classical differential geometry of surfaces. Another consequence of the matrix condition is that the directional derivatives of $\mathbf{x}$ in all directions and at all points are nonzero.

We should note that the standard parametrization of the sphere by the spherical coordinate map $\mathbf{X}(\theta,\phi) = (\cos\theta\sin\phi, \sin\theta\sin\phi, \cos\phi)$ is not quite a regular parametrization for the entire sphere, for we have

$$\frac{\partial \mathbf{X}}{\partial \theta} \times \frac{\partial \mathbf{X}}{\partial \phi} = -\left(\cos\theta\sin^2\phi,\ \sin\theta\sin^2\phi, \cos\phi\sin\phi\right)$$

and this vector vanishes when $\sin\phi = 0$; in other words, when ϕ is an integral multiple of π, or equivalently at the points which map to the north and south poles. — On the other hand, it is possible to find a regular parametrization for the entire sphere such that the domain is a connected region, and an example is given in the file `plane2sphere.pdf`, but this also has an important disadvantage; namely, the associated map $\mathbf{X}$ is very far from being 1–1, either everywhere or "almost everywhere."

Normal thickenings of surfaces

The following result is not always mentioned in differential geometry texts, but it will be helpful for our purposes.

NORMAL THICKENING PRINCIPLE. *Let $\mathbf{x}$ be a regular smooth surface parametrization of class r as above, let*

$$\mathbf{y}(s,t) = \frac{\partial \mathbf{x}}{\partial u}(s,t) \times \frac{\partial \mathbf{x}}{\partial v}(s,t)$$

for $(s,t) \in U$, and let $\Phi(s,t,w) = \mathbf{x}(s,t) + w\,\mathbf{y}(s,t)$ for $(s,t) \in U$ and $w \in (-h,h)$ for some small $h > 0$. Then for each (s,t) there is an $\varepsilon > 0$ (depending on (s,t)) such that the following conclusions hold on the disk

$$D = \{(x,y,z) \in \mathbb{R}^3 \mid (x-s)^2 + (y-t)^2 + z^2 < \varepsilon^2\} :$$

(i) The restriction of Φ to D is $1-1$ and its image is a connected domain V.

(ii) *There is a $\mathcal{C}^r$ inverse map Ψ from V to some connected domain $U_0 \subset U$ containing $(s.t, 0)$ such that $\Psi(\,\Phi(x,y,z)\,) = (x,y,z)$ on U_0.*$\blacksquare$

The map Φ may be viewed as a thickening of $\mathbf{x}$ such that the vertical line segments (s_0, t_0, w) — where the first two variables are held constant — are mapped to curves that are in some sense perpendicular (or **normal**) to the surface at the point $\mathbf{x}(s_0, t_0)$. The figure in the file `thickening.pdf` depicts the special case where the surface is a fragment of a sphere.

Proof. By the Inverse Function Theorem it suffices to show that $D\Phi(s,t,0)$ is invertible for all $(s.t) \in U$, or equivalently that the Jacobian of Φ at these points is always nonzero.

Let $\mathbf{x}_u$ and $\mathbf{x}_v$ denote the partial derivatives of $\mathbf{x}$ with respect to the first and second variables respectively. Then the Jacobian of Φ at $(s,t,0)$ is equal to the value of the vector triple product

$$\left[\mathbf{x}_u,\, \mathbf{x}_v,\, \mathbf{x}_u \times \mathbf{x}_v\right]$$

at (s,t). But the triple product is equal to $|\,\mathbf{x}_u \times \mathbf{x}_v\,|^2$; as noted above, since $D\mathbf{x}$ has rank 2 its columns — which are $\mathbf{x}_u$ and $\mathbf{x}_v$ — are linearly independent, so that the cross product $\mathbf{x}_u \times \mathbf{x}_v$ is nonzero for all $(s,t) \in U$, and therefore its length is positive for all such points. Therefore the Jacobian of Φ is positive at all points $(s,t,0)$ such that $(s,t) \in U$.$\blacksquare$

EXAMPLE. Consider the parametric surface describing a part of the sphere by the spherical coordinate map Σ described above where both θ and ϕ are assumed to lie in $(-\pi, \pi)$. The image of this function is the set of all points on S^2 except for the great circle arc through $(-1,1,0)$ joining the north and south poles. Direct calculation then shows that $\Sigma_u \times \Sigma_v$ is equal to $\sin\phi \cdot \Sigma$. Therefore the normal extension is given by the formula

$$\Phi(\theta, \phi, w) \quad = \quad (1 + w\,\sin\theta) \cdot \Sigma(\theta, \phi)\ .$$

Note that this function maps the entire surface given by the graph $w = -1/\sin\theta$ into $\mathbf{0}$, and therefore the normal extension is not globally 1–1. Furthermore, the Jacobian at points on the curve must vanish because the second partial derivative of Φ at such points is equal to zero (note that the second partial is equal to $(1 + w\,\sin\theta) \cdot \Sigma_2$).

In this example one still knows that there is some $h > 0$ such that Φ is 1–1 and has nonvanishing Jacobian for all (s,t,w) such $|w| < h$ and $(s,t) \in U$. However, it is also possible to construct examples for which one cannot find a positive constant h that works for every point in U. The best one can do in general is find a positive valued continuous function $h(s,t)$ such that Φ is 1–1 and has nonvanishing Jacobian for all (s,t,w) such $|w| < h(s,t)$ and $(s,t) \in U$.

We now proceed to define a concept of surface that is equivalent to the definition on page 126 of O'NEILL (and also the definition in DO CARMO).

Definition. A *geometric regular smooth surface* Σ is a subset of $\mathbb{R}^3$ such that for each $\mathbf{p} \in \Sigma$ there is a smooth 1-1 map ψ defined on some open disk centered at $\mathbf{0}$ in $\mathbb{R}^3$ such that the following hold:

(i) The map ψ sends $\mathbf{0}$ to $\mathbf{p}$, its Jacobian is nowhere zero, and its image W is an open connected domain containing $\mathbf{p}$.

(ii) If r is the radius of the disk on which ψ is defined, then the set $W \cap \Sigma$ is the set of all points of the form $\psi(u,v,0)$ where $u^2 + v^2 < r^2$.

CONSEQUENCE 1. *If $\mathbf{X}$ denotes the restriction of ψ to the set of points whose third coordinate is zero, then $\mathbf{X}$ is a regular smooth parametrization for $\Sigma \cap W$.*

Proof. Let D be the open disk, let D_0 be the corresponding disk in $\mathbb{R}^2$ consisting of all points in D whose third coordinate is equal to zero, and let j denote the inclusion of D_0 in D. Then by the Chain Rule we have that $D\mathbf{X}(u,v) = D\psi(u,v,0) \cdot Dj(u,v)$. Now Dj is simply the 3×2 matrix whose columns are the first two unit vectors, and accordingly it has rank 2, and by hypothesis we know that $D\psi(u,v,0)$ has rank 3. Therefore the composite, which is $D\mathbf{X}(u,v)$, must have rank 2.■

We shall sometimes say that the maps satisfying (i) and (ii) are *thickened (regular smooth) parametrizations* near $\mathbf{p}$.

It is natural to ask why we do not simply define a geometric regular smooth surface to be the image of a smooth 1–1 regular parametrization. The reason for the more complicated definition is to eliminate some "bad" examples that are described at the end of this section.

CONSEQUENCE 2. *If Σ is a above and U is a connected domain such that $\Sigma \cap U$ is not empty, then the latter is also a geometric regular smooth surface. Conversely, if $\Sigma \subset \mathbb{R}^3$ and for each $\mathbf{p} \in \Sigma$ there is an open disk $V_{\mathbf{p}}$ centered at $\mathbf{p}$ such that $\Sigma \cap V_{\mathbf{p}}$ is a geometric regular smooth surface, then Σ itself is a geometric regular smooth surface.*

Proof. We begin by verifying the first inclusion. Let $\mathbf{p}$ be a point in the intersection, let ψ be the map given in the definition above, and let D be the disk on which ψ is defined. The continuity of ψ implies that there is some smaller disk $D' \subset D$ centered at the origin such that the image of D' is contained in U. If we define ψ' to be the restriction of ψ to U, then this restriction satisfies the condition of property (ii) in the definition.

For the second conclusion, if ψ is a map satisfying all the required conditions with respect to $\Sigma \cap V_{\mathbf{p}}$, then it also satisfies these conditions with respect to Σ itself. Since every point $\mathbf{p}$ on the surface lies in a suitable connected domain $V_{\mathbf{p}}$, it follows that property (ii) in the definition of a geometric regular smooth surface is satisfied at every point.■

The basic examples

Before proceeding further we should check that most or all the objects informally described as surfaces are indeed surfaces in the sense of our definition. There are several separate cases to consider.

GRAPHS OF SMOOTH FUNCTIONS. Suppose that we are given a function f that is defined on a connected domain $U \subset \mathbb{R}^2$ and has continuous partial derivatives at every point. Then the graph of f is given by the standard regular smooth parametrization

$$\mathbf{g}(x,y) \quad = \quad \big(x, y,\, f(x,y)\,\big)$$

and we claim that $D\mathbf{g}$ always has rank 2 (or equivalently that the cross product of the first and second partial derivatives of $\mathbf{g}$ is nonzero at all points). Direct computation shows that

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{pmatrix}$$

and it follows that the cross product of the columns has a third coordinate which is equal to $+1$. This cross product will be used repeatedly throughout the remainder of the course, so we shall write it down explicitly:

$$\frac{\partial \mathbf{g}}{\partial x} \times \frac{\partial \mathbf{g}}{\partial y} \quad = \quad \begin{pmatrix} -\frac{\partial f}{\partial x} \\ \\ -\frac{\partial f}{\partial y} \\ \\ 1 \end{pmatrix}$$

The preceding shows that we have a 1–1 regular smooth parametrization for the graph of f. We also need to show that property (ii) in the definition of a geometric regular smooth surface is satisfied. The first step in doing so is to define a 3-dimensional thickening of the parametrization map that is similar to the normal extension discussed previously. Specifically, if W is the connected domain on which f is defined, then we thicken if to a map $\mathbf{F}$ defined on $W \times \mathbb{R}$ by the simple formula

$$\mathbf{F}(u, v, t) \quad = \quad (u,\, v,\, t + f(u, v)\,)\ .$$

It follows immediately that $\mathbf{F}$ is a smooth map with a smooth inverse given by

$$\mathbf{G}(u, v, t) \quad = \quad (u,\, v,\, t - f(u, v)\,)$$

and that the graph of f is the image of $W \times \{0\}$. Suppose now that $\mathbf{p}$ is a point on the graph of f and that $\mathbf{p} = (u,\, v,\, f(u, v)\,)$ for suitable u and v. Let $\mathbf{q}$ denote the vector (u, v), and suppose that $r > 0$ is chosen so that the open 2-dimensional disk of radius r centered at $\mathbf{q}$ lies in W. If D represents the 3-dimensional disk of radius r centered at $\mathbf{0}$ then the necessary map ψ for the point $\mathbf{p}$ is given by $\psi(\mathbf{x}) = \mathbf{F}(\mathbf{x} + \mathbf{q})$; the right hand side is always defined because $\mathbf{x} + \mathbf{q}$ always lies in $W \times \mathbb{R}$ when $\mathbf{x} \in D$.■

In the preceding discussion, we have described graphs in which x and y are the independent variables and z is the dependent variables. Needless to say, one can permute the roles of the three coordinates to consider graphs where each coordinate becomes the dependent variable, and similar considerations show that such subsets are surfaces.■

Notation. Parametrizations of surfaces as graphs of smooth functions are often called *Monge parametrizations* or *Monge patches* in the literature (the pronunciation of the name "Monge" sounds something like "mawzh").

LEVEL SETS OF REGULAR VALUES OF SMOOTH FUNCTIONS. These can be viewed as generalizations of graphs, and they also include the usual quadric surfaces in $\mathbb{R}^3$, at least if one removes a relatively small number of "bad" point that are generally described as singularities; perhaps the simplest example involves the cone defined by the equation $x^2 + y^2 - z^2 = 0$, whose vertex at $\mathbf{0}$ is clearly an exceptional point.

Suppose that we are given a smooth function f defined on a connected domain $U \subset \mathbb{R}^3$, and let C be a constant. We generally expect that the level set defined by the equation $f(x, y, z) = C$ (where (x, y, z) is assumed to lie in U) should define a surface. Perhaps the most fundamental examples of this sort are planes that have equations of the form

$$A\,x \ + \ B\,y \ + \ C\,z \quad = \quad D$$

(where not all of A, B, C are zero) and spheres defined by equations of the form

$$(x - a)^2 \ + \ (y - b)^2 \ + \ (z - c)^2 \quad = \quad r^2$$

(where $r > 0$). The best way to avoid pathologies is to require that C be a **regular value** in the sense that the gradient $\nabla f(x, y, z)$ is not equal to $\mathbf{0}$ if $f(x, y, z) = C$. In both of the cases described above one can check this out directly. For the plane, the gradient is equal to (A, B, C) and this vector is nonzero because we assumed that at least one of the three coefficients was nonzero. In the case of the sphere, the gradient of f at an arbitrary point (x, y, z) is equal to

$$2\left(x - a,\ y - b,\ z - c\right)$$

and therefore vanishes only at the point (a, b, c) which does not lie on the sphere (we assumed that $r > 0$).

We now explain why such level sets are geometric regular surfaces in the sense described above; if we modify our original function by subtracting off the constant C, we obtain a new function such that the gradient is nonzero where the value of the function is zero, so there is no real loss of generality in assuming that $C = 0$. Suppose that $\mathbf{p} = (a, b, c)$ is a point for which $f(a, b, c) = 0$. Since we know that $\nabla f(a, b, c) \neq \mathbf{0}$, at least one partial derivative of f at (a, b, c) is nonzero. If, say, the third partial is nonzero,, then the Implicit Function Theorem implies that there is a small connected domain of the form $V \times W$ containing $\mathbf{p}$ — where V is a connected domain in $\mathbb{R}^2$ containing (a, b) and W is an open interval in $\mathbb{R}$ containing c — and a smooth implicit function g defined on W such that the intersection of the zero set of f with $V \times W$ is equal to the graph of g. We can then use the standard parametrization of a graph as the regular smooth parametrization that is required at the point $\mathbf{p}$. If one of the other partial derivatives at (a, b, c) is zero — say the one with respect to the i^{th} variable — then the the same considerations show that locally the zero set is given by the graph of a function expressing the i^{th} coordinate as a function of the other two.

One can check that this also works for the other basic types of quadric surfaces in the list below, where all exceptional points are noted.

- **Ellipsoids** of the form
$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$
where $a,\ b,\ c \neq 0$. As in the case of the sphere, the gradient of the function on the left hand side vanishes only at $\mathbf{0}$ and the latter does not belong to the level set described above.

- **Hyperboloids** of the form
$$\frac{x^2}{a^2} \pm \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$$
where $a,\ b,\ c \neq 0$. As in the previous case, the gradient of the function on the left hand side vanishes only at $\mathbf{0}$ and the latter does not belong to the level set described above.

- **Cones** of the form
$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - z^2 = 0$$
where $a,\ b \neq 0$ and we restrict to the open connected domain of points that are not equal to $\mathbf{0}$. As in the previous cases, the gradient of the function on the left hand side vanishes only at $\mathbf{0}$ and the latter has been excluded.

- **Elliptic and hyperbolic paraboloids** of the form
$$\frac{x^2}{a^2} \pm \frac{y^2}{b^2} = z$$

where a, $b \neq 0$. In these cases the gradient for the difference of the left and right hand sides never vanishes.

- **Circular, elliptic and hyperbolic cylinders** of the form

$$\frac{x^2}{a^2} \pm \frac{y^2}{b^2} = 1$$

where a, $b \neq 0$. In these previous case, the gradient for the difference of the left and right hand sides vanishes only at points where $x = y = 0$, and no point of the form $(0, 0, z)$ belongs to one of the level sets described above.

- **Parabolic cylinders** of the form

$$\frac{x^2}{a^2} = z$$

where $a \neq 0$. In these cases the gradient for the difference of the left and right hand sides never vanishes.

This list is not quite exhaustive, but the only types of nondegenerate quadrics that are missing are given by two planes that either intersect in a line (the hyperbolic cylinder equation with the right hand side set equal to 0 rather than 1) and pairs of parallel lines defined by an equation of the form $x^2 = a^2 > 0$ (see the end of Section IV.3 for more information on this point). In the first case one must exclude the entire z-axis, but in the second case it is not necessary to exclude any points at all.

CYLINDRICAL SURFACES. We have already discussed some standard examples of cylindrical surfaces. Generalizations of these examples turn out to play an important role in many aspects of geometry, so it is worthwhile to explain how some of them can be parametrized. The simples examples of cylindrical surfaces arise when one takes a curve in $\mathbb{R}^2$ defined by $y = f(x)$ and considers the set of all points $(x, y, z) \in \mathbb{R}^3$ such that $y = f(x)$. If J is the interval upon which f is defined, then this surface is the subset of $J \times \mathbb{R} \times \mathbb{R}$ consisting of all points satisfying the equation $y - f(x) = 0$, so this set will be a geometric surface because the gradient of $y - f(x)$ is the nonzero vector $(-f'(x),\, 1,\, 0)$. In this case one also has a simple explicit parametrization

$$\mathbf{x}(u, v) = (u,\, f(u),\, v)$$

that maps $J \times \mathbb{R}$ to the surface in a 1–1 onto fashion.

In the preceding example, one uses lines that are perpendicular to the xy-plane, but one can also form such surfaces using a family of mutually parallel lines such that these lines are neither parallel to nor contained in the xy-plane. The corresponding smooth parametrization in such cases is given by the formula

$$\Sigma(t, s) = (t,\, f(t),\, 0) + s \cdot (a, b, c)$$

where $c \neq 0$.

SURFACES OF REVOLUTION. Several of the quadric surfaces described above can be viewed as surfaces of revolution about a coordinate axis, and more general surfaces of revolution also play an important role in geometry. Therefore we shall consider the two basic types of examples that one encounters in single variable calculus courses. Given a curve $y = f(x)$ as above such that $f(x) > 0$ for all x, then we can construct a corresponding surface of revolution in $\mathbb{R}^3$ about the x-axis. Such a surface is defined by an equation of the form $y^2 + z^2 = f(x)^2$ on the set $J \times \mathbb{R} \times \mathbb{R}$, where J is an open interval on which f is defined, and an explicit 1–1 global parametrization is given by

$$\Sigma(t, \theta) \quad = \quad \big(t,\, f(t)\, \cos\theta,\, f(t),\, \sin\theta\big)\ .$$

Verification that this description yields a geometric surface is left to the reader as an exercise.

Similarly, if we are given a curve $y = f(x)$ as above that is defined on an interval for which x is always positive, then we can also construct a corresponding surface of revolution in $\mathbb{R}^3$ about the y-axis. In this case an explicit 1–1 global parametrization is given by

$$\Sigma(t, \theta) \quad = \quad \big(t\, \cos\theta,\, f(t),\, t\, \sin\theta\big)\ .$$

Alternatively, one can view a surface of revolution about the y-axis as given by the equation $y = f(\sqrt{x^2 + z^2}\,)$; if f is defined on the interval (a, b) where $a > 0$, then the domain of definition for the corresponding function of x and z is the annulus defined by the inequalities

$$a^2 \quad < \quad x^2 + z^2 \quad < \quad b^2\ .$$

We shall give a slight generalization of this which shows that the torus given by rotating a circle such as $(x-1)^2 + y^2 = 1$ about the y-axis is a surface in the sense of these notes. Suppose we are given a simple closed curve $\mathbf{x}$ in $\mathbb{R}^2$ which can also be described as the set of solutions to $F(u, v) = 0$ where $\nabla F(a, b) \neq \mathbf{0}$ at all points such that $F(a, b) = 0$, and suppose that the first coordinates of all solutions to $F(u, v) = 0$ are greater than some positive number a. A parametrization of the resulting surface of revolution is given by

$$\mathbf{X}(t,\, \theta) \quad = \quad \big(u(t)\, \cos\theta,\, v(t),\, u(t)\, \sin\theta\big)$$

and if we set $G(x, y, z) = F(\sqrt{x^2 + z^2},\, y)$, then the surface of revolution consists of all points such that $G(x, y, z) = 0$. In order to verify that this defines a surface in our sense, we need to show that the gradient of G is nonzero at all points of the zero set of G. Here is a sketch of the proof: At each point (u, v) such that $F(u, v)$ we know that either the first partial derivative $F_1(u, v)$ or the second partial derivative $F_2(u, v)$ is nonzero. Suppose now that $G(x, y, z) = 0$ and let $u = \sqrt{x^2 + z^2}$ and $v = y$. If the second partial derivative of F is nonzero at (u, v), then the second partial derivative of G is also nonzero at (x, y, z). If the first partial derivative of F is nonzero at (u, v) and $x \neq 0$, then elementary calculations show that the first partial derivative of G is also nonzero at (x, y, z), while if the first partial derivative is nonzero and $z \neq 0$, then the third partial derivative of G is also nonzero at (x, y, z). Since $u > a > 0$ by hypothesis we know that $x^2 + z^2 > a^2 > 0$, and therefore at least one of x and z is always zero; this proves that the gradient of G is nonzero at every point of the zero set.

RULED SURFACES. More generally, one can define another important generalization of cylindrical surfaces that also includes the cone that are **ruled** in the sense that one has parametrizations for the entire surface of the form

$$\mathbf{X}(u, v) \quad = \quad \mathbf{a}(u) + v \cdot \mathbf{b}(u)$$

where $\mathbf{a}'(u)$ is never zero and the vectors $\mathbf{a}'(u)$ and $\mathbf{b}(u)$ are always linearly independent. Here are some basic examples that are not cylindrical in the sense described above:

- A **hyperbolic paraboloid.** Consider the surface of this type defined by the equation $z = x^2 - y^2$. The right hand side factors as a product $(x - y)\,(x + y)$, so the intersection of the surface with the plane $x - y = C$ is just the line at which the planes $x - y = C$ and $z = C(x + y)$ intersect. This leads to the definition of parameters $u = x - y$ and $v = x + y$, and one can use these to parametrize the surface as

$$\mathbf{X}(u, v) \quad = \quad (\tfrac{1}{2}(u + v),\ \tfrac{1}{2}(u - v),\ u\,v\,)\ .$$

  Here the curves defined by holding either u or v constant are straight lines, and one can rewrite the parametrization in the form $\mathbf{y}(u) \ + \ v\,\mathbf{g}(u)$ where

$$\mathbf{y}(u) \quad = \quad \tfrac{1}{2}\,u\,(\mathbf{e}_1 + \mathbf{e}_2)$$

  and

$$\mathbf{g}(u) \quad = \quad \tfrac{1}{2}\,(\mathbf{e}_1 + \mathbf{e}_2) \ + \ u\,\mathbf{e}_3\ .$$

- A **hyperboloid of one sheet.** Consider the surface of this type defined by the equation $x^2 + y^2 - z^2 = 1$. One can check directly that this surface can be parametrized using the function

$$(\cos u,\ \sin u\,) \quad + \quad v \cdot (-\sin u,\ \cos u, 1)$$

  and that $\mathbf{a}(u) = (\cos u,\ \sin u)$ and $\mathbf{b}(u) = (-\sin u,\ \cos u, 1)$ satisfy the basic conditions described above.

- A **cone.** We shall only consider the nonsingular piece of the cone $x^2 + y^2 - z^2 = 0$ in the upper half plane where $z > 0$. In this case the parametrization is given by

$$\mathbf{X}(u, v) \quad = \quad (v\cos u,\ v\sin u,\ v)$$

  where $u \in \mathbb{R}$ and $v > 0$. One can give ruled parametric equations by the alternate formulas

$$(\cos u,\ \sin u,\ 1) + v \cdot (\cos u,\ \sin u,\ 1)$$

  where again $u \in \mathbb{R}$ but this time $v > -1$.

- The **Möbius strip.** Intuitively, this is formed by taking a rectangle $A\,B\,C\,D$ for which the length $|AB| = |CD|$ is much greater than the width $|BC| = |AD|$ and gluing sides BC and AD so that B corresponds to D and A corresponds to C. One can model this using the parametric equations

$$\mathbf{X}(u, v) \quad = \quad (\cos u,\ \sin u, 0) \ + \ v \cdot \Big(\cos u \cos(u/2),\ \sin u \cos(u/2),\ \sin(u/2)\Big)$$

  where $u \in \mathbb{R}$ and $v \in (-\frac{1}{2}, \frac{1}{2})$ (or one can take $|v| < \varepsilon$ for some arbitrary ε that is positive but less than 1).

In order to show this satisfies the condition for a surface, it will suffice to find a set of open domains U_i such that every point in the image of the parametrization $\mathbf{X}$ lies in one of the domains U_1 and that on each set U_i the intersection of the Möbius strip with the zero set of some well behaved smooth function on U_i. Geometrically, the key to doing this is to look at the intersection of the surface with the planes containing the z-axis, which are defined in cylindrical coordinates by equations of the form $\theta = C$. In such planes one sees that the points of the Möbius strip are the points satisfying $(r-1)^2 + z^2 < \varepsilon^2$ and either $z = (1-r)\tan\frac{1}{2}C$ if C is not an odd multiple of π or else by $1 - r = z\cot\frac{1}{2}C$ if C is not an even multiple of 2π. Therefore, on the set of points in $\mathbb{R}^3$ satisfying $(r-1)^2 + z^2 < \varepsilon^2$ and **either** $x > 0$ **or** $y \neq 0$, the intersection with the Möbius strip is given by the equation $z = (1-r)\tan\frac{1}{2}\theta$, while on the set of points satisfying $(r-1)^2 + z^2 < \varepsilon^2$ and **either** $x < 0$ **or** $y \neq 0$, the intersection with the Möbius strip is given by the equation $(1-r) = z\cot\frac{1}{2}\theta$.■

Here are some online references, including some with animations showing the one-sidedness of the Möbius strip.

`http://www.worldofescher.com/gallery/A29.html`
`http://www.mikejwilson.com/solidworks/`(*continue with next line*)
`files/mobius_II_animation.zip` (*This requires RealOne Player*.)
`http://www.physlink.com/Education/AskExperts/ae401.cfm`
`http://www.uta.edu/optics/sudduth/4d/`(*continue with next line*)
`nonorientable/moebius_strip/math/mathematics.htm`
`http://www.mapleapps.com/categories/animations/gallery/anim_pg3.shtml`
`http://www.tattva.com/vladi/director.html#6` (*Scroll down the Movie List to the last entry, which is called "Mobius strip." There are QuickTime and RealOne Player versions of this loop*.)
`http://mathworld.wolfram.com/MoebiusStrip.html` (*This is a curious animation*.)

Significant counterexamples

On the basis of our examples thus far, it is natural to ask whether the image of a parametrized surface is always a geometric surface. It turns out that the answer is negative, even if one restricts attentions to simple parametrizations that are globally 1–1. Here is one counterexample: Consider the figure 8 curve $\varphi(t) = (\sin 2t, \sin t)$ for $t \in (0,\, 2\pi)$. One then has an associated cylindrical surface with regular smooth parametrization $\Sigma(t,w) = (\sin 2t,\, \sin t,\, w)$ for $t \in (0, 2\pi)$ and $w \in \mathbb{R}$. This parametrization is also 1–1, but its image fails to satisfy the definition of a geometric surface when $\mathbf{p} = \mathbf{0}$. The key to seeing this is the following simple observation:

PROPOSITION. *Let Σ be a geometric regular smooth surface in $\mathbb{R}^3$, and let $\mathbf{p} \in \Sigma$. Define $\mathbf{K_p}$ to be the set of all vectors in $\mathbb{R}^3$ that are realizable as tangent vectors $\mathbf{y}'(0)$, where $\mathbf{y}$ is a smooth curve entirely contained in Σ such that $\mathbf{y}(0) = \mathbf{p}$. Then $\mathbf{K_p}$ is a 2-dimensional vector subspace of $\mathbb{R}^3$.*

Proof. Let ψ be a smooth 1-1 map ψ defined on some open disk centered at $\mathbf{0}$ in $\mathbb{R}^3$ such that (i) it sends $\mathbf{0}$ to $\mathbf{p}$, its Jacobian is nowhere zero, and its image W is an open connected domain containing $\mathbf{p}$, (ii) if r is the radius of the disk on which ψ is defined, then the set $W \cap \Sigma$ is the set of all points of the form $\psi(u, v, 0)$ where $u^2 + v^2 < r^2$.

Let φ be the inverse mapping to ψ, and suppose that $\mathbf{y}$ is a curve of the type described in the conclusion of the proposition. By restricting to a small interval centered at 0, we may as well assume that the image of $\mathbf{y}$ is contained in the image of ψ so that $\phi \circ \mathbf{y}$ is defined. This is a curve in the uv-plane, so its tangent vector at 0 also lies in this plane. By the Chain Rule, the tangent vector to $\mathbf{y} = \psi \circ (\varphi \circ \mathbf{y})$ lies in the subspace of $\mathbb{R}^3$ spanned by $D\psi(\mathbf{0})\mathbf{e}_1$ and $D\psi(\mathbf{0})\mathbf{e}_2$. Conversely, every vector in this subspace is the tangent vector of a curve in the surface of the form $\psi(t\,\mathbf{v})$ where $\mathbf{v}$ lies in the subspace of $\mathbb{R}^3$ spanned by the first two unit vectors.■

Returning to the example, we now consider all curves of the form

$$\big(\sin 2a\,t,\ \sin a\,(t - c\,\pi), b\,t\,\big)$$

where a and b are arbitrary real numbers and $c = 0$ or 1. Each of these curves lies entirely in the image of the parametrized surface, and at parameter value t each curve passes through $\mathbf{0}$. What are the tangent vectors to these curves? They are equal to $(2\,a, \pm\,a, b)$. We claim there is no 2-dimensional vector subspace W of $\mathbb{R}^3$ that contains this set. To see this, note that the set of all tangent vectors described above contains the 2-dimensional subspace W_0 spanned by $(2,1,0)$ and $(0,0,1)$, and if W is a 2-dimensional subspace containing these and possibly other tangent vectors, then $W = W_0$. On the other hand, the given set of tangent vectors includes $(2. - 1.0)$, which is definitely not in W_0. — It follows that the image of the 1–1 parametrization map is not a geometric regular smooth surface in this case.■

ANOTHER (more complicated) EXAMPLE. The cylindrical surface in Exercise 19 on pages 68–69 of do Carmo illustrates another way in which the image of a 1–1 parametrization may fail to be a smooth surface. According to the defining conditions, for every point $\mathbf{p}$ of a geometric surface Σ, for every connected domain W containing $\mathbf{p}$ there is a connected subdomain $U \subset W$ containing $\mathbf{p}$ such that every other point in $\Sigma \cap U$ can be joined to $\mathbf{p}$ by a smooth curve lying entirely in $\Sigma \cap U$. This property fails to hold for the surface described in the exercise; specifically, consider the disk W of radius $\frac{1}{4}$ about the origin and the points $\mathbf{q}_n$ with coordinates

$$\left(\frac{1}{n\,\pi}, 0\right) \ .$$

We claim that there are no smooth curves in $\Sigma \cap W$ joining the origin to such points. If there were, then by the Intermediate Value Theorem for each value of t between 0 and $1/n\,\pi$ there would be points on these curves, and hence on the surface Σ, whose first coordinates are equal to t. However, examination of the graph of $\sin(1/x)$ shows that the only point with first coordinate $2/(\,(2n+1)\,\pi)$ on this curve have second coordinates with absolute values ≥ 1 and therefore such points do not lie in W. If U is an arbitrary connected domain containing the origin, then it contains a disk of some positive radius, and this disk contains all but finitely many of the points $\mathbf{q}_n$. Since one cannot join these points to $\mathbf{0}$ in $\Sigma \cap W$ by smooth curves lying completely within the latter intersection, one certainly cannot find such curves in the even smaller intersection $\Sigma \cap U$. Therefore Σ does not satisfy the second condition required for a geometric surface.

Piecewise smooth surfaces

In Unit I we noted that there are many contexts in which it is necessary to consider curves that are piecewise smooth, and likewise there are many contexts in which it is necessary to consider piecewise smooth surfaces. In particular, such objects play important roles in multivariable integral calculus, and perhaps the most obvious examples are given by the surfaces of cubes and cylinders. For the sake of completeness, we note that formal definitions of piecewise smooth surfaces are given in Section 7.1 of Colley (see Definition 1.3 on page 413) and in the following online file:

`http://math.ucr.edu/~res/math10B/comments0701.pdf`

III.3 : Tangent planes

(Lipschutz, Chapter 8)

Special cases of tangent planes are introduced in multivariable calculus courses, particularly for surfaces that are graphs of functions with continuous partial derivatives. In order to specify a plane, it is enough to specify a point on the plane and a line that is perpendicular — or **normal** — to that plane; the latter can be given by vector that determines the perpendicular direction. For graphs, the point is supposed to have the form $(x,\ y,\ f(x,y)\,)$, and the the direction vector is equal to

$$\left(-\frac{\partial f}{\partial x},\ -\frac{\partial f}{\partial y},\ 1\right)$$

which we have seen before in another context. Accordingly, the first degree equation defining the tangent plane at $(a,b,,f(a,b)\,)$ is given by

$$z - f(a,b) \quad = \quad f_x(a,b)\cdot(x-a)\ +\ f_y(a,b)\cdot(y-b)$$

where f_x and f_y denote the partial derivatives with respect to x and y respectively.

There is an important characterization of tangent planes in terms of tangent lines.

PROPOSITION. *If $\mathbf{x}$ is a regular smooth curve in the graph of a smooth function f, and $\mathbf{x}(0) = (a,b,f(a,b)\,)$, then the tangent line to $\mathbf{x}$ at parameter value $t=0$ lies in the tangent plane. Conversely, if L is a line through $(a,b,f(a,b)\,)$ that lies in the tangent plane, then there is a regular smooth curve $\mathbf{x}$ in the graph of f such that $\mathbf{x}(0) = (a,b,f(a,b)\,)$ and the tangent line to the curve at $(a,b,f(a,b)\,)$ is L.*

Proof. Suppose that $\mathbf{x}$ is a regular smooth curve with parametric equations given by

$$\mathbf{x}(t) \quad = \quad \big(\,u(t),\ v(t),\ w(t)\,\big)\ .$$

Then the relation $w = f(u,v)$ and the chain rule imply that $w'(0) = f_u(a,b)\cdot u'(0)\ +\ f_v(a,b)\cdot v'(0)$, and it follows immediately by substitution that the tangent line to $\mathbf{x}$ at parameter value 0 lies in the tangent plane to the graph at $(a,b,f(a,b)\,)$.

Conversely, every line L of the given type has a parametrization of the form

$$(a,b,f(a,b)\,) + t\cdot(M,\ N,\ P\,)$$

where $-M\,f_x(a,b) - N\,f_y(a,b) + P = 0$. Choose $r > 0$ so that the open disk of radius r is contained in the domain U on which f is defined. If we let

$$r_0 \quad = \quad \min\left\{\frac{r}{|M|+1},\ \frac{r}{|N|+1}\right\}$$

then for $|t| < r_0$ the parametrized segment $(a + t\,M,\ b + t\,N)$ lies in U, and the curve

$$\mathbf{x}(t) \quad = \quad \big(\,a + t\,M,\ b + t\,N,\ f(a + t\,M,\ b + t\,N)\,\big)$$

lies on the graph of f. Furthermore, we know that

$$\mathbf{x}'(0) \quad = \quad \big(\,M, N, f_u(a,b)M + f_v(a,b)N\,\big)$$

and by the first sentence of this paragraph the third coordinate is equal to P. Therefore the tangent line to $\mathbf{x}$ at parameter value $t = 0$ is equal to L.■

One can also define tangent planes for regular parametrizations by a similar formula. Specifically, if $\mathbf{X}$ is a parametrization for the surface that is defined on the connected domain U and $(a, b) \in U$, then the tangent plane at parameter value (a, b) is the unique plane through $\mathbf{X}(a, b)$ whose normal direction is given by

$$\frac{\partial \mathbf{X}}{\partial u}(a, b) \times \frac{\partial \mathbf{X}}{\partial v}(a, b) \ .$$

If $\mathbf{X}$ is a graph parametrization with z given as a function $f(x, y)$, then the the cross product above reduces to the familiar vector

$$\begin{pmatrix} -f_x(a, b) \\ -f_y(a, b) \\ 1 \end{pmatrix}$$

and therefore the definition of tangent plane for parametrizations reduces to the previous definition if $\mathbf{X}$ is a graph parametrization.

The previous characterization of tangent planes generalizes as follows: If L is a line through $\mathbf{X}(a, b)$ in the tangent plane, then every direction for vector for L is perpendicular to the cross product of $\mathbf{x}_u(a, b)$ and $\mathbf{x}_v(a, b)$ and hence is a linear combination of these two vectors; for the sake of definiteness, express a direction vector for L in the form $M\,\mathbf{x}_u(a, b) + N\,\mathbf{x}_v(a, b)$. It follows that the curve $\mathbf{y}(t) = \mathbf{X}(a + t\,M,\, b + t\,N)$ has tangent vector $\mathbf{y}'(0) = M\,\mathbf{x}_u(a, b) + N\,\mathbf{x}_v(a, b)$. Thus L is the tangent line to a curve through $\mathbf{X}(a, b)$ that lies in the image of the parametrized surface. Conversely, if we are given a curve in the image of $\mathbf{X}$, whose value at $t = 0$ is equal to (a, b), by the Inverse Function Theorem we know that for $|t|$ sufficiently small we may write the curve as

$$\mathbf{y}(t) \quad = \quad \mathbf{X}\big(u(t),\, v(t)\big)$$

and therefore we have

$$\mathbf{y}'(0) \quad = \quad \frac{\partial \mathbf{X}}{\partial u}(a, b) \cdot u'(0) \ + \ \frac{\partial \mathbf{X}}{\partial v}(a, b) \cdot v'(0) \ .$$

Since this vector is perpendicular to the normal direction for the tangent plane, it follows that the tangent line to $\mathbf{y}$ at parameter value $t = 0$ lies in the tangent plane.■

The tangent planes described above may be described as all vectors of the form $\mathbf{p} + \mathbf{w}$, where $\mathbf{w}$ is the tangent vector to a curve that goes through $\mathbf{p}$ and lies completely in the parametrized surface. If P is an arbitrary plane containing the point $\mathbf{p}$ and its normal direction is $\mathbf{N}$, then the set of all vectors having the form $\mathbf{y} - \mathbf{p}$ is merely the set of all vectors that are perpendicular to $\mathbf{N}$, and hence they form a 2-dimensional subspace of $\mathbb{R}^2$ that we shall call the *space of tangent vectors at* $\mathbf{p}$ *for the parametrization of the surface*. By construction this subspace is either equal to the tangent plane at $\mathbf{p}$ or else it is the unique plane through the origin that is parallel to the tangent space; the first holds if $\mathbf{0}$ lies in the tangent plane, and the second holds if it does not.

ALTERNATE CHARACTERIZATION OF TANGENT PLANES. *The tangent plane to the parametrized surface* $\mathbf{X}$ *at parameter value* (a, b) *is the unique plane through* $\mathbf{p} = \mathbf{X}(a, b)$ *that is parallel or equal to the* 2*-dimensional subspace spanned by* $[D\,\mathbf{X}(a, b)]\mathbf{e}_i$ *for* $i = 1, 2$.

This is essentially contained in earlier results, the point being that the direction vectors for lines L in the tangent plane containing $\mathbf{p}$ all have the form $[D\,\mathbf{X}(a, b)]\mathbf{v}_i$, where $\mathbf{v}$ is a linear combination of $\mathbf{e}_1$ and $\mathbf{e}_2$.■

SPECIALIZATION TO LEVEL SETS. Suppose we have a surface that is defined as the set of all solutions to the equation $f(x,y,z) = 0$, where f is a smooth function such that $\nabla f(x,y,z) \neq \mathbf{0}$ whenever $f(x,y,z) = 0$. The following result provides a very simple description of the normal direction to the tangent plane.

GRADIENTS ARE THE NORMALS TO LEVEL SETS. *Let f be as above, and suppose that $f(a,b,c) = 0$. Then there is a local parametrization of the surface near (a,b,c) such that the normal direction for the tangent plane at (a,b,c) is equal to $\nabla f(a,b,c)$.*

Proof. In principle, it suffices to do this when the third coordinate of $\nabla f(a,b,c)$ is nonzero; the other cases follow by interchanging the roles of the three coordinates.

If the third coordinate is zero, then there is a small connected domain V containing (a,b,c) such that the set of solutions for $f(a,b,c) = 0$ is given by the graph of some smooth function $z = g(u,v)$. Therefore the normal direction of the plane at (a,b,c) is given by the familiar vector $(-g_u(a,b), -g_v(a,b), 1)$. On the other hand, the Implicit Function Theorem implies that $g_u = -f_u/f_z$ and $g_v = -f_v/f_z$, and therefore the gradient is equal to the scalar product of the partial derivative $f_z(a,b,c)$ with $(-g_u(a,b), -g_v(a,b), 1)$.■

IMPORTANT SPECIAL CASE. For the sphere defined by the equation $x^2 + y^2 + z^2 - r^2 = 0$, the gradient of $f(x,y,z) = x^2 + y^2 + z^2 - r^2$ is equal to $2\,(x,y,z)$, and this confirms a well known property for the tangent planes to points on a sphere: *They are perpendicular to the radial line at the point of contact.*

This preceding result describes the tangent plane in a manner that is independent of the choice of parametrization; in particular, if all three coordinates of $\nabla f(a,b,c)$ are nonzero, then one gets three distinct parametrizations locally by viewing each coordinate as the graph of a function in the other two near (a,b,c). For an arbitrary geometric regular smooth surface Σ, it is natural to expect that **all** regular local smooth parametrizations for the surface near a point $\mathbf{p}$ yield the same tangent plane at $\mathbf{p}$. The following result proves this is always the case.

COMPATIBILITY THEOREM. *Let Σ be a geometric regular smooth surface, let $\mathbf{p} \in \Sigma$, and let ψ_1 and ψ_2 be thickened regular smooth parametrizations at $\mathbf{p}$. Let $\mathbf{Q}$ be the subspace of $\mathbb{R}^3$ spanned by the first two unit vectors. Then the images of $\mathbf{Q}$ under the maps $D\psi_1(\mathbf{0})$ and $D\psi_2(\mathbf{0})$ are equal.*

It follows that the common image is the natural candidate for the 2-dimensional space of tangent vectors to Σ at $\mathbf{p}$.

Proof. Suppose that ψ_i is defined on an open disk $\mathbf{D}(r_i)$ of radius $r_i > 0$ centered at $\mathbf{0}$. By the continuity of the mappings ψ_i and their inverses, we can find a real number $s_2 > 0$ such that $s_2 < r_2$ and ψ_2 maps the open disk $\mathbf{D}(s_2)$ into $\psi_1(\,\mathbf{D}(r_1)\,)$. It follows that there is a smooth map

$$G : \mathbf{D}(s_2) \to \mathbf{D}(r_1)$$

defined by $G(\mathbf{w}) = \psi_1^{-1}(\,\psi_2(\mathbf{w}\,)$. By construction it follows that $\psi_1 \circ G = \psi_2$. Furthermore, by the conditions on thickened parametrizations we know that the Jacobian of G is always nonzero, $G(\mathbf{0}) = \mathbf{0}$, and

$$G(u,v,0) \quad = \quad (\,x(u,v),\, y(u,v),\, 0)$$

for suitable smooth functions x and y. The last formula shows that if $\mathbf{q}$ lies in $\mathbf{Q}$, then $[D\,G(\mathbf{0})](\mathbf{q})$ also lies in $\mathbf{Q}$; the converse also holds because $[D\,G(\mathbf{0})](\mathbf{q})$ is invertible (hence the image of $\mathbf{Q}$ is a 2-dimensional subspace that we know is contained in $\mathbf{Q}$, and therefore it must be equal to $\mathbf{Q}$ — since the derivative is 1–1 nothing else can map into $\mathbf{Q}$).

If we apply the Chain Rule to $\psi_1 \circ G = \psi_2$, it follows that

$$D\psi_1(\mathbf{0}) \cdot D\,G(\mathbf{0}) = D\psi_2(\mathbf{0}) \ .$$

Let $\mathbf{q}$ be an arbitrary vector in the subspace $\mathbf{Q}$ spanned by the first two unit vectors as above. Since we have seen that $\mathbf{q} \in \mathbf{Q}$ implies $[D\,G(\mathbf{0})](\mathbf{q}) \in \mathbf{Q}$, it follows that $D\psi_2(\mathbf{0})\mathbf{q}$ lies in the image of $\mathbf{Q}$ under $D\psi_1(\mathbf{0})$. Conversely, suppose that we are given a vector of the form $D\psi_1(\mathbf{0})\mathbf{p}$ for some $\mathbf{p} \in \mathbf{Q}$. Then by the preceding paragraph we may write $\mathbf{p} = [D\,G(\mathbf{0})](\mathbf{q})$ for some $\mathbf{q}$ in $\mathbf{Q}$, and by the formula displayed at the beginning of this paragraph it follows that $D\psi_1(\mathbf{0})\mathbf{p} = D\psi_2(\mathbf{0})\mathbf{q}$. Therefore the two subspaces in question are equal as required.■

In fact, we have the following characterization of tangent planes to geometric surfaces which can be stated without using local parametrizations.

COROLLARY. *Let Σ be a geometric regular smooth surface, and let $\mathbf{p} \in \Sigma$. Then the tangent space to Σ at $\mathbf{p}$ is the set of all vectors of the form $\mathbf{p} + \gamma'(0)$ where $\gamma(t)$ is a smooth curve such that its image lies entirely in Σ and $\gamma(0) = \mathbf{p}$.*■

This corollary can be also derived without appealing directly to the Compatibility Property, for the Normal Thickening Principle implies that implies that the space described in the corollary is equal to the image of the uv-plane under the maps $D\mathbf{X}(u_0, v_0)$ and $D\Phi(u_0, v_0, 0)$, where $\mathbf{X}(u, v) = \Phi(u_0, v_0, 0) = \mathbf{p}$.

III.4 : The First Fundamental Form

(Lipschutz, Chapter 9)

The First and Second Fundamental Forms are comparable to the curvature and torsion of a curve in that surfaces are locally characterized up to geometric congruence by these forms just as curves are so characterized by their curvatures and torsions. The two fundamental forms are also important for numerous other reasons as well. In particular, **the First Fundamental Form is crucial to virtually all work in the differential geometry of surfaces and their higher dimensional generalizations.**

There are two definitions of the fundamental form, one for parametrizations and one for geometric surfaces. We shall begin with the latter and then indicate how it is given in terms of parametrizations.

The definitions of the First and Second Fundamental Forms for a geometric surface both involve an object that is generally called the *tangent space* in differential geometry.

Definition. Let S be a geometric surface in $\mathbb{R}^3$, and for each $\mathbf{p} \in S$ let $T_{\mathbf{p}}(S)$ denote the 2-dimensional vector space of tangent vectors to S at $\mathbf{p}$; in the previous section we showed that this 2-dimensional subspace did not depend upon the choice of local parametrization. The **tangent space** of S, denoted by $\mathbf{T}(S)$, is the defined to be the set

$$\{\,(\mathbf{p}, \mathbf{q}) \in \mathbb{R}^3 \times \mathbb{R}^3 \mid \mathbf{p} \in S \ \text{ and } \ \mathbf{q} \in T_{\mathbf{p}}(S)\,\}$$

In some sense this consists of all the tangent planes to points in $\mathbb{R}^3$, but we have spread things out over six dimensions so that the analogs of tangent planes at different points do not have any vectors in common (in contrast, note that every point on the unit sphere $x^2 + y^2 + z^2 = 1$ lies on more than one tangent plane; in elementary plane geometry, this corresponds to showing that there are two tangents to a circle going through a given external point). Projection onto the first factor defines a map τ_S from $\mathbf{T}(S)$ to S, which corresponds geometrically to sending each tangent vector to the "point of application." Similarly, one can view projection Φ onto the last three coordinates as defining a map from $\mathbf{T}(M)$ to $\mathbb{R}^3$ that sends a tangent vector to its associated "free vector" (no point of application) in $\mathbb{R}^3$.

EXAMPLES. If S is the (regular) level set of of zeros for some smooth function $f(x, y, z)$, then $\mathbf{T}(S)$ is the set of all points

$$\big(x, y, z,\ X, Y, Z\big) \quad \in \quad \mathbb{R}^3 \times \mathbb{R}^3$$

such that $f(x, y, z) = 0$ and (X, Y, Z) is perpendicular to $\nabla f(x, y, z)$. If we specialize further to a sphere defined by $x^2 + y^2 + z^2 - r^2 = 0$ we see that the tangent space consists of all 6-tuples such that (x, y, z) lies on the sphere and is perpendicular to (X, Y, Z).

Definition. Let $\mathbf{T}^{(2)}(M)$ be the set of all ordered pairs of points $(\mathbf{v}_1,\ \mathbf{v}_2)$ in $\mathbf{T}(M) \times \mathbf{T}(M)$ such that $\tau_S(\mathbf{v}_1) = \tau_S(\mathbf{v}_2)$. The *First Fundamental Form* of S is the map $\mathbf{I}_S$ ending $(\mathbf{v}_1,\ \mathbf{v}_2)$ to the usual inner product $\langle\, \Phi(\mathbf{v}_1),\ \Phi(\mathbf{v}_2)\,\rangle$ of two vectors in $\mathbb{R}^3$.

Perhaps the simplest motivation for the First Fundamental Form is that it can be used to describe arc lengths. In particular, if $\mathbf{x}$ is a parametrized smooth curve lying entirely on S and we define a tangent lifting $TL(\mathbf{x})$ of $\mathbf{x}$ to $\mathbf{T}(M)$ by the formula

$$TL(\mathbf{x})(u) \quad = \quad \big(\,\mathbf{x}(u),\ \mathbf{x}'(u)\,\big)$$

then the length of the curve is given by

$$\int_a^b \Big(\mathbf{I}_S(TL\mathbf{x}(t),\ TL\mathbf{x}(t)\Big)^{1/2}\, dt\ .$$

In fact, this formula motivates the definition of the First Fundamental Form for parametrized surfaces as follows:

Definition. Let $\mathbf{X}$ be a regular smooth surface parametrization defined on some connected domain U. Then the *First Fundamental Form* of $\mathbf{X}$ is the function defined on $U \times \mathbb{R}^2 \times \mathbb{R}^2$ by the formula

$$\mathbf{I}_{\mathbf{X}}(\mathbf{p}; \mathbf{y},\ \mathbf{z}\,) \quad = \quad \langle [D\mathbf{X}(u)](\mathbf{y}),\ [D\mathbf{X}(u)](\mathbf{z})\,\rangle$$

where the right hand side denotes the inner product of two vectors in $\mathbb{R}^3$.

It follows immediately that if we have a curve $\mathbf{c}$ defined in U, then the length of $\mathbf{X} \circ \mathbf{c}$ can be computed either by means of the first fundamental form as defined here or by the previous definition of the first fundamental form. For the sake of completeness, if the curve $\mathbf{c}$ is given in parametric form by $\mathbf{c}(t) = (u(t), v(t)\,)$, then by the Chain Rule then the tangent vectors to $\mathbf{X} \circ \mathbf{c}$ are equal to

$$\frac{\partial \mathbf{X}}{\partial u}\,\frac{d\,u}{d\,t} \ + \ \frac{\partial \mathbf{X}}{\partial v}\,\frac{d\,v}{d\,t}$$

and the length of the curve is given by to the following integral:

$$\int_a^b \ \left|\frac{\partial \mathbf{X}}{\partial u}\,\frac{d\,u}{d\,t} \ + \ \frac{\partial \mathbf{X}}{\partial v}\,\frac{d\,v}{d\,t}\right|\, dt$$

Classical references use somewhat different notation that we shall now describe. Consider the square of the expression inside the length integral given above. Using the bilinear nature of the inner product we may write this as follows:

$$\left(\left|\frac{\partial \mathbf{X}}{\partial u}\right|^2 \left(\frac{du}{dt}\right)^2 \; + \; 2\left(\frac{\partial \mathbf{X}}{\partial u}\cdot\frac{\partial \mathbf{X}}{\partial v}\right) \frac{du}{dt}\,\frac{dv}{dt} \; + \; \left|\frac{\partial \mathbf{X}}{\partial v}\right|^2 \left(\frac{du}{dt}\right)^2 \right) dt\,dt$$

Using the standard formal convention of setting

$$dw \quad = \quad \frac{dw}{dt}\,dt$$

we may rewrite this expression in the form

$$E(u,v)\,du\,du \; + \; 2\,F(u,v)\,du\,dv \; + \; G(u,v)\,dv\,dv$$

where the smooth functions E, F and G are defined by

$$E \;=\; \frac{\partial \mathbf{X}}{\partial u}\cdot\frac{\partial \mathbf{X}}{\partial u} \qquad F \;=\; \frac{\partial \mathbf{X}}{\partial u}\cdot\frac{\partial \mathbf{X}}{\partial v} \qquad G \;=\; \frac{\partial \mathbf{X}}{\partial v}\cdot\frac{\partial \mathbf{X}}{\partial v}$$

This is the classical formula for the First Fundamental Form.

Digression: Diagonalizing symmetric matrices

The First Fundamental Form of a surface in $\mathbb{R}^3$ is essentially a special case of a geometrical structure called a **Riemannian metric**. Before describing such objects formally, we shall briefly review some important facts about symmetric square matrices which are closely tied to the definition of a Riemannian metric. These will play an extremely important role throughout the rest of this course.

PROPOSITION. *If A is a symmetric $n \times n$ matrix and X, Y are $n \times 1$ column vectors then we have*

$$AX \cdot Y \quad = \quad AY \cdot X$$

where "$\cdot$" denotes the dot product.

Sketch of proof. Direct calculation and the symmetric matrix identity $a_{i,j} = a_{j,i}$ show that both of the displayed expressions are equal to $\sum\, a_{i,j} x_i x_j$.■

THEOREM. *If A is a 2×2 symmetric matrix over the real numbers, then A has an orthonormal basis of eigenvectors.*

In our setting, an eignevector for A is a nonzero 2×1 column matrix X such that $AX = cX$ for some real number c; as usual, c is called an **eigenvalue** for A.

Proof. We recall an important fact from linear algebra: *For a given real number c, there is a nonzero vector X such that $AX = cX$ if and only if $A - cI$ is not invertible, or equivalently if and only if c is a root of the characteristic polynomial $\chi_A(t) = \det(A - tI)$.*

The symmetric matrix A has the form

$$\begin{pmatrix} a & d \\ d & b \end{pmatrix}$$

for suitable real numbers a, b, d. We claim that the roots of the characteristic polynomial $\chi_A(t)$ are real. Direct calculuation shows that

$$\chi_A(t) \quad = \quad t^2 \ - \ (\mathrm{trace} A)t \ + \ (\det A) \quad = \quad t^2 \ - \ (a+b)t \ + \ (ab-d^2)$$

and thus the roots of this polynomial are real if and only if the discriminant

$$(a+b)^2 \ - \ 4(ab-d^2)$$

is nonnegative. But direct calculation shows that the latter equals

$$a^2 \ - \ 2ab \ + \ b^2 \ \ + d^2 \quad = \quad (a-b)^2 \ + \ d^2$$

and the latter is nonnegative because it is a sum of squares.

The final step in the proof is to verify that A has an orthonormal basis of eigenvectors. There are two cases depending upon whether the characteristic polynomial has one or two roots. If there is only one root then the discriminant is zero, which means that $(a-b)^2 + d^2 = 0$, and the latter holds if and only if $a - b = d = 0$. This implies that $A = cI$ for $c = a = b$, and the existence of an orthonormal basis of eigenvectors is obvious because every nonzero vector is an eigenvector.

Assume now that the characteristic polynomial has two distinct roots c_1 and c_2, and for $i = 1, 2$ let X_i be a nonzero vector such that $AX_i = c_i X_i$. If we apply the identity in the preceding result, we obtain the following:

$$AX_1 \cdot X_2 \quad = \quad c_1 X_1 \cdot X_2 \ , \qquad AX_1 \cdot X_2 \quad = \quad X_1 \cdot AX_2 \quad = \quad X_1 \cdot c_2 X_2 \quad = \quad c_2 X_1 \cdot X_2$$

Combining these equations, we see that $c_1(X_1 \cdot X_2) = c_2(X_1 \cdot X_2)$. Since $c_1 \neq c_2$ it follows that $X_1 \cdot X_2 = 0$, so that the nonzero vectors X_1 and X_2 are orthogonal. If we multiply these vectors by the reciprocals of their lengths we obtain a pair of orthonormal vectors in the space of 2×1 column matrices which are also eigenvectors. Since the space of 2×1 column matrices is 2-dimensional and nonzero orthogonal vectors are linearly independent, it follows that our orthonormal set is a basis.■

Abstract Riemannian metrics

In the middle of the nineteenth century G. F. B. Riemann observed that certain generalizations of the First Fundamental Form had were strongly connected to other central problems in geometry including the subject of Noneuclidean Geometry. In simplified form, his insight was to consider arbitrary expressions of the form

$$g(u,v) \quad = \quad E(u,v)\, du\, du \ + \ 2\, F(u,v)\, du\, dv \ + \ G(u,v)\, dv\, dv$$

where E, F and G are smooth functions on some connected domain U such that the real symmetric matrix

$$\mathbf{M}(u,v) \quad = \quad \begin{pmatrix} E(u,v) & F(u,v) \\ F(u,v) & G(u,v) \end{pmatrix} \quad = \quad \begin{pmatrix} g_{1,1}(u,v) & g_{1,2}(u,v) \\ g_{2,1}(u,v) & g_{2,2}(u,v) \end{pmatrix}$$

(where $g_{1,2} = g_{2,1}$ in the second matrix) is said to be *positive definite* in one of the following equivalent senses:

(1) For every nonzero vector $\mathbf{x}$ the inner product $\langle \mathbf{M}(u,v)\mathbf{x},\, \mathbf{x} \rangle$ is positive.

(2) The eigenvalues of $\mathbf{M}(u,v)$ are all positive real numbers.

(3) The diagonal entries and determinant of $\mathbf{M}(u,v)$ are all positive.

This type of structure is called a **riemannian metric**.

Given a riemannian metric defined on a connected domain U and a regular smooth curve $\mathbf{x}(t) = (\, u(t),\, v(t)\,)$ in U, then one can define the *length* of $\mathbf{x}$ with respect to this riemannian metric by the formula

$$\int_a^b \sqrt{\langle \mathbf{M}\big(u(t), v(t)\,\big)\mathbf{x}'(t),\, \mathbf{x}'(t)\rangle}\; dt$$

because positive definiteness implies that the expression inside the square root sign is always positive. The classical Noneuclidean Geometry developed by Bólyai, Lobachevsky and others can then be described by taking U to be the open unit disk about the origin in $\mathbb{R}^2$ and the riemannian metric equal to the so-called Poincaré metric:

$$\frac{dx\,dx \;+\; dy\,dy}{(1-x^2-y^2)^2}$$

To illustrate these ideas, we shall compute the Poincaré length of the closed segment $[0, r]$ on the x-axis, where $0 < r < 1$. The formulas imply that the length is given by the integral

$$\int_0^r \frac{1}{1-t^2}\,dt \quad = \quad \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \ .$$

Notice that this expression goes to $+\infty$ as $r \to 1$.

Shortest curves joining two points. In this and other systems involving riemannian metrics, one basic question is to determine the shortest smooth, or piecewise smooth, curve joining two points. For the Poincaré metric there are two cases.

(I) If one has points $\mathbf{x}$ and $\mathbf{y}$ in U such that the line joining them contains the origin, then the shortest curve is the ordinary line segment joining them. However, the length of this curve with respect to the Poincaré metric will **NOT** be equal to its Euclidean length.

(II) If $\mathbf{0}$ is not on the line joining $\mathbf{x}$ and $\mathbf{y}$, then the shortest curve is a circular arc whose endpoints are $\mathbf{x}$ and $\mathbf{y}$, where the circle K containing the arc meets the unit circle $x^2+y^2 = 1$ orthogonally; *i.e.*, for each of the two points where K and the unit circle meet, the tangent lines to K and the unit circle at the common point are perpendicular to each other. Proving this is definitely not a trivial matter and requires methods beyond the scope of this course.

Here are some online references regarding Noneuclidean Geometry:

`http://mathworld.wolfram.com/PoincareHyperbolicDisk.html`

`http://mathworld.wolfram.com/HyperbolicGeometry.html`

Incidentally, relativity theory uses a generalization of riemannian metric in which the positive definiteness condition is replaced by something weaker. Perhaps the most basic example is the Lorentz metric given by

$$dt\,dt \;-\; dx\,dx \;-\; dy\,dy \;-\; dz\,dz \ .$$

Riemannian metrics are basic examples of a common theme in more advanced studies of differential geometry; the main idea is summarized in the following passage from the Preface to O'Neill:

> Every surface has a differential and integral calculus of its own, strictly analogous to the familiar calculus of the plane.

For example, at every point of a surface one has a 2-dimensional vector space of tangent vectors, and one can devise a meaningful definition of smooth mapping between two smooth surfaces. Some aspects of this principle will play a significant role in later units of these notes.

III.5 : Surface area

(Lipschutz, Chapter 9)

This is mainly a review of material covered in multivariable calculus courses. Two textbook references are to Sections 13.5 and 14.5 on pages 971–977 and 1051–1060 of *Calculus* (Seventh Edition), by Larson, Hostetler and Edwards, and also Section 6.3 on pages 382–395 of *Basic Multivariable Calculus*, by Marsden, Tromba and Weinstein. Here are a couple of online references which also cover the topic:

`http://math.ucr.edu/~res/math10B/comments0701.pdf`

`http://math.ucr.edu/~res/math10B/discoballs.pdf`

The basic idea behind surface area formulas is to find approximations using areas of pieces of various tangent planes. For example, suppose we have the graph of a function $z = f(x, y)$ and we want to compute the area of the portion of the surface lying over some rectangle in the plane whose sides lie on lines that are parallel or equal to the coordinate axes. One first cuts the large rectangle into many smaller rectangles, then chooses a point (x, y) in each rectangle, and next for each point one finds the area of the portion of the tangent plane $(x, y, f(x, y))$ which lies above the small rectangle containing the original point (x, y), and finally one adds up all these areas to get an approximation to the surface area we wish to compute. If we take increasingly larger decompositions into smaller and smaller rectangles and let the maximum lengths and widths go to zero, the one expects the limit to be the surface area, and this is indeed the case. A more detailed discussion of this appears on pages 306–307 of O'Neill (see also Section 2–8 of do Carmo; most multivariable calculus texts also discuss this topic at some length).

Important cautionary note. In view of the standard description of arc length of a "reasonable" curve Γ as the limit of broken line curves that are inscribed in Γ, it is natural to ask is surface area could be defined more simply by considering polyhedral pieces that are inscribed in surface and defining the area of the surface to be the limit of the areas of such polyhedral approximations. However, **this approach does not always yield the expected answer, even in simple cases like the lateral portion of the cylinder defined by** $x^2 + y^2 = 1$ and $0 \leq z \leq 1$. A discussion of this issue, including some pictures, is contained in the file `surface-area-critique.pdf`, which is taken from the 1946 edition of Widder, *Advanced Calculus* (a Second Edition also exists); the specific reference is Subsection 7.4.

Standard special cases. For a surface parametrization given as the graph of a smooth function f, the area of the portion of the surface over a reasonable subset A in the plane is given by the integral

$$\int_A \sqrt{1 + f_1(x, y)^2 + f_2(x, y)^2}\, dx\, dy$$

where f_1 and f_2 denote the partial derivatives with respect to the first and second variables. If we are given a regular 1–1 surface parametrization $\mathbf{X}$ and A is a reasonable subset of the connected domain U on which $\mathbf{X}$ is defined, then the standard formula for the area is given by

$$\int\int_A |\mathbf{X}_u \times \mathbf{X}_v| \, du \, dv$$

where $\mathbf{X}_u$ and $\mathbf{X}_v$ denote the partial derivatives of $\mathbf{X}$. The area can also be expressed in terms of the coefficients of the First Fundamental Form as follows:

$$\text{Area} \quad = \quad \int_A \sqrt{E\,G - F^2} \, du \, dv$$

Derivation. This follows directly from the standard length formula

$$|\mathbf{X}_u \times \mathbf{X}_v|^2 \quad = \quad |\mathbf{X}_u|^2 \cdot |\mathbf{X}_v|^2 \; - \; |\mathbf{X}_u \cdot \mathbf{X}_v|^2$$

and the definitions of the functions E, F and G in the preceding section of these notes.∎

The preceding discussion shows how to find the areas of portions of a surface but it does not directly address the question of finding the area of the entire surface. In order to do this, one needs to decompose the surface into disjoint or nonoverlapping pieces, find the areas of the different pieces separately, and then add the results together. In many cases one can also simplify the computations by using parametrizations that are well behaved *almost everywhere*; making this term precise mathematically is beyond the scope of this course, but some simple examples include cases where the bad behavior is limited to some finite set of points or some finite collection of regular smooth curves. For example, if one wants to compute the surface area of the unit sphere, one can take the spherical coordinate parametrization defined for $\theta \in [0, 2\pi]$ and $\phi \in [0, \pi]$. This parametrization is not 1–1 on boundary points and $\mathbf{X}_\theta \times \mathbf{X}_\phi$ vanishes at some boundary points, but it is a regular smooth 1–1 parametrization away from these boundary points and thus gives the area for, say, the portion of the sphere not including the semicircular meridian through the north and south pole and the point $(1, 0, 0)$. The meridian by itself has zero area, and this is why there is no problem using the formula even though things do not work well on the boundary.

Areas associated to riemannian metric. If we are given a riemannian metric g defined on an open region U in $\mathbb{R}^2$, we can use the First Fundamental Form to define areas for a reasonable subset A of U by generalizing the formula derived above:

$$\text{Area}(A) \quad = \quad \int\int_A \sqrt{g_{1,1}\, g_{2,2} - g_{1,2}^2} \, du \, dv$$

In particular, if U is the open unit disk and g is the previously introduced Poincaré metric for noneuclidean geometry

$$\frac{dx\,dx \; + \; dy\,dy}{(1 - x^2 - y^2)^2}$$

we can compute the hyperbolic area of the closed disk A_r defined by $x^2 + y^2 \le r^2 < 1$ as follows: The coefficients $g_{i,j}$ are given by

$$g_{1,1} \quad = \quad g_{2,2} \quad = \quad \frac{1}{(1 - x^2 - y^2)^2} \; , \qquad g_{1,2} \quad = \quad g_{2,1} \quad = \quad 0$$

and hence the area of A_r is given by

$$\int\int_{A_r} \frac{1}{(1 - x^2 - y^2)^2} \, dx \, dy \; .$$

We can compute this by converting to polar coordinates, obtaining the equivalent integral

$$\int_0^{2\pi} \int_0^r \frac{t}{(1-t^2)^2}\, dt\, d\theta \quad = \quad 2\pi \left. \frac{1}{2(1-t^2)} \right|_0^r \quad = \quad \pi \, \frac{r^2}{1-r^2} \ .$$

Notice that this goes to ∞ as $r \to 1$.

III.6 : Curves as surface intersections

(No suitable text reference)

Given two distinct planes in $\mathbb{R}^3$ that have at least one point in common, a standard axiom or theorem in 3-dimensional Euclidean geometry states that their intersection is a line. Specifically, if the point $\mathbf{a}$ lies on both planes $\mathbf{P}$ and $\mathbf{Q}$ and normal vectors to these planes are given by $\mathbf{p}$ and $\mathbf{q}$ respectively, then the line in question consists of all vectors expressible as a sum $\mathbf{a}+t\,(\mathbf{p}\times\mathbf{q})$, where t is some real number; examples are discused on page 755 of Larson, Hostetler and Edwards). There are many familiar situations in which the intersection of more general surfaces are also curves, and some of these will play a key role in the definition of curvature for surfaces. Therefore we shall spend some time discussing the realizations of curves as intersections of surfaces.

If Σ is a sphere and $\mathbf{p}$ is a point on Σ, then for almost every plane $\mathbf{Q}$ passing through $\mathbf{p}$ the intersection of Σ and $\mathbf{Q}$ is a circle, the only exception being when $\mathbf{Q}$ is the tangent plane at$\mathbf{p}$.

Consider next the intersection of a sphere whose radius is $b > 0$ and whose center is the origin with a cylinder $\mathbf{H}$ whose axis is the z-axis. If the radius a of this cylinder is less than b, then the intersection consists of the two circles with parametric equations

$$(a \cos t,\ a \sin t,\ \pm\sqrt{b^2-a^2}\,)$$

which are the latitude lines on Σ that lie $\cos^{-1}(a/b)$ radians above the equatorial circle formed by the intersection of Σ with the xy-plane. The point of this example is that the intersection is not one curve but two curves, and it is meant to suggest that in general we should first consider the intersection of two surfaces locally. In fact, we shall generally restrict attention to the local situation.

Returning to the intersection of a sphere and a plane, or the intersection of two distinct planes, elementary calculations show that the normal lines for two such surfaces at points of intersection are always distinct (except when one has the tangent plane to a point on a sphere). Furthermore, the same thing happens at the intersection points of the sphere and cylinder that were discussed above. Alll these examples serve as motivation for the following general result, which shows that the intersection of two level surfaces Σ_1 and Σ_2 is locally a curve near a point **provided** the tangent planes to Σ_1 and Σ_2 and the common points are distinct.

TRANSVERSE INTERSECTIONS OF LEVEL SURFACES. *Let f and g be smooth functions defined on a connected domain U, let $\Sigma(f)$ and $\Sigma(g)$ denote their zero sets, and suppose that ∇f and ∇g are nonzero at all points of $\Sigma(f)$ and $\Sigma(g)$ respectively. Suppose that $\mathbf{p}$ lies on $\Sigma(f)\cap\Sigma(g)$ and that $\nabla f(\mathbf{p})$ and $\nabla g(\mathbf{p})$ are linearly independent (i.e., the intersection is* **transverse** *at $\mathbf{p}$). Then there is an open domain U containing $\mathbf{p}$ such that $U\cap\Sigma(f)\cap\Sigma(g)$ is a regular smooth curve.*

Another example. Consider the surfaces of revolution formed by rotating the standard circle $x^2 + y^2 = 4$ and ellipse

$$\frac{x^2}{9} \;+\; (y-1)^2 = 1$$

about the y-axis. The intersection of these surfaces splits into two pieces, one of which consists of the point $(0, 2, 0)$ and the other of which is the circle parametrized by

$$\left(\frac{\sqrt{63}}{4}\,\cos\theta,\; \frac{1}{4},\; \frac{\sqrt{63}}{4}\,\sin\theta \right) \; .$$

At points of the latter the tangent planes to the two surfaces are distinct, but at $(0, 2, 0)$ they are not. This illustrates that the intersection of two surfaces might be transverse at some points but not necessarily at others.

Proof of transverse intersection property. This will be a consequence of the Implicit Function Theorem. Let

$$\mathbf{H}(x,y,z) \quad = \quad \big(\, f(x,y,z),\; g(x,y,z)\,\big)$$

so that $\mathbf{H}$ is a smooth function and $D\,\mathbf{H}$ is the 2×3 matrix whose rows are the gradients of f and g. Since the gradients are linearly independent at $\mathbf{p}$, it follows that $D\,\mathbf{H}(\mathbf{p})$ has rank 2. Therefore there is a 2×2 submatrix of $D\,\mathbf{H}(\mathbf{p})$ whose determinant is nonzero. It will suffice to consider the case where the determinant of the square submatrix constructed from the last two columns is nonzero; the other cases can be handles similarly by interchanging the roles of the variables.

Express $\mathbf{p}$ in coordinates as (a, b, c). We then know there is an open interval U_0 containing a and a smooth 2-dimensional vector valued function k on U_0 such that $k(a) = (b, c)$ and for all $x \in U_0$ and (y, z) close to (b, c), say in some connected domain V_0 containing (b, c) we have $\mathbf{H}(x, y, z) = 0$ if and only if $(y, z) = k(x)$. It follows that the intersection of the surfaces with $U_0 \times V_0$, which is just the intersection of the latter with the zero set of $\mathbf{H}$, is equal to the image of the regular parametrized curve whose first coordinate is given by t and whose second and third coordinates are given by $k(t)$.$\blacksquare$

Note. One can describe the tangent line to this curve at $\mathbf{p}$ in terms of f and g; specifically, it is the line through $\mathbf{p}$ whose direction is given by $\nabla f(\mathbf{p}) \times \nabla g(\mathbf{p})$. This follows because the tangent vector at $\mathbf{p}$ is perpendicular to both gradients.$\blacksquare$

COMPLEMENT. *The same result holds for arbitrary surfaces Σ_1 and Σ_2 provided the tangent planes at a common point $\mathbf{p}$ are distinct.*

The proof of this depends upon the following observation.

LEMMA. *If Σ is a geometric surface and $\mathbf{p} \in \Sigma$, then there is a connected domain U containing $\mathbf{p}$ and a smooth real valued function $f : U \to \mathbb{R}$ such that the gradient of f is nonzero at all points in the zero set of f, and this zero set is equal to $\Sigma \cap U$.*

Proof of Lemma. By the definition of a geometric surface there is a smooth 1-1 map ψ defined on some open disk centered at $\mathbf{0}$ in $\mathbb{R}^3$ such that (i) the map ψ sends $\mathbf{0}$ to $\mathbf{p}$, its Jacobian is nowhere zero, and its image W is an open connected domain containing $\mathbf{p}$, (ii) if r is the radius of the disk on which ψ is defined, then the set $W \cap \Sigma$ is the set of all points of the form $\psi(u, v, 0)$ where $u^2 + v^2 < r^2$. Let φ be the inverse to ψ, and let c_3 be the smooth map on $\mathbb{R}^3$ which sends each point to its third coordinate. Then the zero set of the function $c_3 \circ \varphi$ is equal to $\Sigma \cap U$, so it is only necessary to verify that the gradient is nonzero at all such points. However, the gradient of this map is given by the third column of the matrix $D\varphi(\mathbf{x})$, and since we know that this matrix is invertible for all $\mathbf{x} \in W$ (by the corresponding fact for $D\psi$), it follows that the gradient is indeed nonzero as required.■

Proof of Complement. Since the conclusion is local, it suffices to take the intersections of the surfaces with some open disk containing $\mathbf{p}$, and by the preceding result we can choose the radius of this disk small enough so that the two surfaces are level sets. Furthermore, the conditions on the tangent planes imply that the gradients of the associated functions must be linearly independent at $\mathbf{p}$. Therefore we may apply the transverse intersection property to show that locally the intersection of the two surfaces is given by a regular smooth curve.■

The preceding results yield the following "intuitively obvious" fact:

COROLLARY. *Let Σ be a geometric surface, let $\mathbf{p} \in \Sigma$, and suppose that $\mathbf{Q}$ is a plane through $\mathbf{p}$ that is not the tangent plane to the surface at $\mathbf{p}$. Then there is a connected domain U containing $\mathbf{p}$ such that $\Sigma \cap \mathbf{Q} \cap U$ is a regular smooth curve through $\mathbf{p}$.*■

Finally, we shall show that every regular smooth curve can be realized locally as the intersection of two surfaces. There are corresponding global statements, but their proofs require more mathematical tools than we currently have or wish to develop in this course.

REALIZATION PRINCIPLE. *Let $\mathbf{x}$ denote a regular smooth curve defined on a closed interval $[-h,h]$such that $\mathbf{x}(0) = \mathbf{p}$. Then there is a connected domain U containing $\mathbf{p}$ and two geometric surfaces Σ_1 and Σ_2 such that $\Sigma_1 \cap \Sigma_2 \cap U$ is equal to the intersection of U with the image of $\mathbf{p}$.*

Proof. A regular smooth curve is locally 1–1, so we can assume that $h > 0$ is so small that $\mathbf{x}$ is globally 1–1 on the interval $[-h,h]$.

Since $\mathbf{x}'(0)$ is nonzero, one can find vectors $\mathbf{y}$ and $\mathbf{z}$ such that $\mathbf{x}'(0)$, $\mathbf{y}$ and $\mathbf{z}$ form a basis for $\mathbb{R}^3$. Consider the smooth map $\mathbf{F}$ defined by

$$F(t,u,v) \quad = \quad \mathbf{x}(t) \ + \ u\,\mathbf{y} \ + \ v\,\mathbf{z} \ .$$

By construction $D\,\mathbf{F}(0,0,0)$ is the matrix whose columns are given by the basis $\mathbf{x}'(0)$, $\mathbf{y}$ and $\mathbf{z}$ and therefore this derivative matrix is invertible. Applying the inverse function theorem, we can find an open disk U centered at $\mathbf{0}$ on which $\mathbf{F}$ has a smooth inverse and nonzero Jacobian; let $r > 0$ be the radius of this disk, where we choose $r < h$. By construction, if L denotes the x-axis, then the image of $L \cap U$ is a piece of the curve $\mathbf{x}$.

We claim that if we shrink the radius sufficiently we can find a subdisk $U_0 \subset U$ such that $\mathbf{F}(U_0)$ does not contain any other points on the curve aside from those that lie in the image of $L \cap U_0$. Consider the images of the closed intervals $[-h,\ -r]$ and $[r,\ h]$. Neither image contains $\mathbf{0}$, and by continuity the distance from points on these curves to $\mathbf{0}$ assumes some positive minimum value, say m. If we take U_0 to be the disk of radius s centered at $\mathbf{0}$, where $0 < s < m$, then it will follow that $\mathbf{F}(U_0 \cap L)$ is equal to the intersection of U_0 with the image of the original curve defined by $\mathbf{x}$.

Finally, if we let Σ_1 and Σ_2 be the images of the intersections of the xy-plane and xz-plane under $\mathbf{F}$ and set $W = \mathbf{F}(U)$, then it follows that Σ_1 and Σ_2 are surfaces and the intersection $\sigma_1 \cap \Sigma_2 \cap U_0$ is just the portion of $\mathbf{x}$ that lies in U_0.■

III.7 : Map projections

(No suitable text reference)

The theory of mapmaking provides many examples of parametrizations for the unit sphere that are 1–1 on the regions where they are defined (and points in these regions are marked to indicate the locations of their images on the sphere). Of course, an important objective of mapmaking is to minimize the distortion that occurs when one maps a flat surface onto a curved one, so the most useful examples have special properties reflecting this goal. For example, in some cases straight line segments in the plane will correspond to great circle arcs on the sphere, in other cases the angle between certain curves in the plane will be equal to the angle between their images on the sphere (for example, if one curve is a latitude or longitude lines), and in still others one wishes to minimize or at least strictly control the distortions in areas that occur when one takes a region in the plane and maps it to a piece of the sphere. One standard and example is the familiar map projection due to G. Mercator (1512–1594); a picture of this projection can be found on page 11 of the document

`http://math.ucr.edu/~res/math153/history08.pdf`

and a more detailed mathematical discussion is given in Section 3.7 of D. V. Widder, *Advanced Calculus*, Second Edition, Dover, New York, 1989, ISBN: 0-486-66103-2). Explicit formulas relating the xy coordinates to the latitude λ and longitude ϕ can be found at the following online site:

`http://mathworld.wolfram.com/MercatorProjection.html`

The Mercator projection is very good near the equator, but it becomes extremely distorted as one approaches the north and south poles. For example, in the Mercator projection Greenland appears to be larger than Australia, but in fact Australia is more than $3\frac{1}{2}$ times as large as Greenland. A drawing to compare the areas of these land masses and further examples of Mercator distortion are given in the following online reference;

`http://en.wikipedia.org/wiki/Mercator_projection`

One way of creating a more accurate map near the poles is to use **stereographic projection**; this mapping method was essentially known to Greek mathematicians like Hipparchus of Rhodes (*c.* 190 B.C.E. – *c.* 120 B.C.E.) more than 2000 years ago and it probably even dates back to ancient Egyptian mathematics. The stereographic projection from the North Pole N is defined at every point P of the sphere except the North Pole itself, and it sends P to the point where the line N and P meets the tangent plane to the South Pole (see `stereographic0.pdf`). The explicit formula for stereographic projection at a point $(x, y, z) \in S^2$ is given by

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

If we write $F(x, y, z) = (u, v)$, then one can solve for (x, y, z) in terms of (u, v); this parametrization is discussed in the exercises.

Finally, many other elementary aspects of mapmaking theory are covered from a mathematical viewpoint in the following recent book, which is written at the undergraduate level:

> T. G. Freeman, *Portraits of the Earth: A Mathematician Looks at Maps*, American Mathematical Society, Providence RI, 2002, ISBN 0–8218–3255–7.

IV. Oriented surfaces

Given a surface Σ and a point $\mathbf{p}$ on Σ, it is meaningful to talk about the normal line to Σ at $\mathbf{p}$ which is simply the unique line that is perpendicular to the tangent plane at $\mathbf{p}$. Each such line may be viewed as having two distinct senses of direction, and an orientation is basically a way of specifying a sense of direction for every normal line to the surface. The theory of surface integrals in multivariable calculus requires the use of orientations, and Stokes' Theorem is a basic result for which orientations of surfaces are absolutely necessary; this is particularly reflected by the fact that the result does not hold for the Möbius strip. Orientations also play an important role in describing the curvature properties of a surface, and curvature was originally defined for surfaces using orientations. Although many basic curvature properties do not depend upon orientations, the original approach to curvature using orientations provides numerous important insights that are often difficult at best to understand from other approaches.

IV.1 : Normal directions and Gauss maps

(Lipschutz, Chapter 9)

If Σ is a surface and $\mathbf{p}$ is a point of Σ, then the space of tangent vectors to Σ at $\mathbf{p}$ is a 2-dimensional subspace of $\mathbb{R}^3$. The orthogonal complement of this subspace is the 1-dimensional space of normal vectors. At each point $\mathbf{p}$ there are precisely two normal vectors that have unit length.

Definition. If Σ is a surface, then an *orientation* of Σ is a continuous map $\mathbf{N} : \Sigma \to \mathbb{R}^3$ (i) such that for each $\mathbf{p} \in \Sigma$ the vector $\mathbf{N}(\mathbf{p})$ is a normal vector to $\mathbf{p}$ with unit length, (ii) for each $\mathbf{p}$ there is an open disk D centered at $\mathbf{p}$ on which $\mathbf{N}$ extends to a smooth map from D to $\mathbb{R}^3$. A surface Σ is *orientable* if one can define an orientation for Σ, and if $\mathbf{N}$ is an orientation for Σ we say that $(\Sigma, \mathbf{N})$ is an oriented surface (or surface with orientation).

Clearly orientations are not unique; in particular, if $\mathbf{N}$ is an orientation for Σ then so is $-\mathbf{N}$. Furthermore, if one considers the pair of parallel planes defined by the equation $z^2 = 1$, then one clearly has at least two choices for the orientation on each plane (namely, take $\mathbf{N} = \pm\,\mathbf{e}_3$, where the signs can be chosen independently on each of the planes). However, the following result shows that locally there are only two possible orientations for a surface.

PROPOSITION. *Let Σ be a surface, let $\mathbf{p}$ be a point of Σ, and let $\mathbf{A}$ and $\mathbf{B}$ be orientations of Σ. Then there is an open disk U containing $\mathbf{p}$ such that $\mathbf{B} = \pm\,\mathbf{A}$ on $U \cap \Sigma$.*

Proof. For each point $\mathbf{q} \in \Sigma$ we know that $\mathbf{B}(\mathbf{q}) = \pm\,\mathbf{A}(\mathbf{q})$. In particular, for each $\mathbf{q}$ this means that either $|\mathbf{B}(\mathbf{q}) \mp \mathbf{A}(\mathbf{q})| = 0$ or else $|\mathbf{B}(\mathbf{q}) \mp \mathbf{A}(\mathbf{q})| = 2$. Suppose that $\mathbf{B}(\mathbf{q}) = \mathbf{A}(\mathbf{q})$. Then by continuity we know there is some small disk D containing $\mathbf{p}$ such that $|\mathbf{B}(\mathbf{q}) \mp \mathbf{A}(\mathbf{q})| < 1$ for all $\mathbf{q}$ in $\Sigma \cap D$. Since there are only two choices for the distance $|\mathbf{B}(\mathbf{q}) \mp \mathbf{A}(\mathbf{q})|$ and one of them is greater than 1, it follows that the distance must be zero on all such points, so that $\mathbf{B}$ is equal to $\mathbf{A}$ on $\Sigma \cap D$. Similarly, if $\mathbf{B}(\mathbf{q}) = -\mathbf{A}(\mathbf{q})$, then there is an open disk V centered at $\mathbf{p}$ such that $\mathbf{B}$ is equal to $-\mathbf{A}$ on $\Sigma \cap D$.■

There are two fundamental examples of orientable surfaces for which orientations are easy to construct.

LEVEL SURFACES. Suppose that Σ is the zero set of a smooth function f defined on a connected domain U, where as usual we assume that ∇f is always nonzero on Σ. In this case we know that the gradient of f is perpendicular to the 2-dimensional space of tangent vectors, and therefore we may define an orientation by the formula

$$\mathbf{N}(\mathbf{p}) \quad = \quad \frac{1}{\nabla f(\mathbf{p})} \cdot \nabla f(\mathbf{p}) \ .$$

The condition that $\mathbf{N}$ extend to a smooth disk about each point is automatically satisfied because one can use the formula to define $\mathbf{N}$ on some open disk containing $\mathbf{p}$ on which the gradient is nonzero.■

SURFACES WITH GOOD GLOBAL PARAMETRIZATIONS. Suppose now that there is a 1–1 regular parametrization $\mathbf{X}$ for Σ that is defined on some connected domain U. Let

$$\Omega(\mathbf{p}) \quad = \quad \frac{\partial \mathbf{X}}{\partial u}(s,t) \ \times \ \frac{\partial \mathbf{X}}{\partial v}(s,t)$$

where $(s,t) \in U$ is the unique point such that $\mathbf{X}(s,t) = \mathbf{p}$, so that $\Omega(\mathbf{p})$ is a nonzero vector that is perpendicular to the tangent plane at $\mathbf{p}$. Then we may define an orientation by the formula

$$\mathbf{N}(\mathbf{p}) \quad = \quad \frac{1}{\Omega(\mathbf{p})} \cdot \Omega(\mathbf{p}) \ .$$

To verify the extension condition, let Φ be the normal thickening described in Section III.2, take

$$\Omega \quad = \quad \frac{\partial \Phi}{\partial u} \ \times \ \frac{\partial \Phi}{\partial v}$$

note that Ω is nonzero on an open disk containing $\mathbf{p}$, and as before take $\mathbf{N}$ to be Ω divided by its length.

A nonorientable example. The standard example is the Möbius strip; the usual way of explaining nonorientability is by studying the behavior of a unit normal vector to the surface as it travels around the central circle in the surface; as one goes around this circle once, the normal vector moves continuously from itself to minus itself. We can make this mathematically precise using the parametrization at the end of Section III.2. The central circle in the Möbius strip is just the circle in the xy-plane with parametric equations $(\cos t, \sin t, 0)$ for $t \in [0,\, 2\pi]$, and the curve with parametric equations

$$\Big(-\,\cos u \sin(u/2),\ -\,\sin u \sin(u/2),\ \cos(u/2)\,\Big)$$

defines a unit normal to the surface at $\mathbf{z}(t)$ for every choice of t. Direct computation shows that $\mathbf{y}(0) = (0,0,1)$ and $\mathbf{y}(2\,\pi) = (0,0,-1)$. In order to prove that the Möbius strip is not orientable, it suffices to show that this sort of thing cannot happen if one has an orientation for a surface.

PROPOSITION. *Let $(\Sigma, \mathbf{N})$ be an oriented surface, let $\mathbf{z}$ be a regular smooth closed curve defined on $[0,\, 2\pi]$ and taking values in Σ, and let $\mathbf{y}$ be another smooth closed curve such that $\mathbf{y}(t)$ is a multiple of $\mathbf{N}(\mathbf{z}(t)\,)$ for all t, with $\mathbf{y}(0) = \mathbf{N}(\mathbf{z}(0)\,)$ and $\mathbf{y}(2\pi) = -\mathbf{N}(\mathbf{z}(2\pi)\,)$. Then there is a point $\alpha \in [0,\, 2\pi]$ such that $\mathbf{y}(\alpha) = \mathbf{0}$.*

Proof. The hypothesis that that $\mathbf{y}(t)$ is a multiple of $\mathbf{N}(\mathbf{z}(t)\,)$ for all t and a standard inner product formula imply that

$$\mathbf{y}(t) \quad = \quad \langle \mathbf{y}(t),\, \mathbf{N}(\mathbf{z}(t)\,)\,\rangle \cdot \mathbf{N}(\mathbf{z}(t)\,)$$

and therefore the coefficient

$$f(t) \quad = \quad \langle \mathbf{y}(t),\, \mathbf{N}(\mathbf{z}(t)\,)\,\rangle$$

is a continuous function of t. By construction we have $f(0) = 1$ and $f(2\pi) = -1$, and therefore the Intermediate Value Theorem implies the existence of some $\alpha \in (0, 2\pi)$ such that $f(\alpha) = 0$.■

The Gauss map

The orientation map for an oriented surface is also known as the *Gauss map*. This map will play an important role in the rest of the course.

It is instructive to look at some examples. First of all, if Σ is a plane, then the normal lines at different points are all parallel to each other, and in fact if one views a plane as the set of points satisfying an equation of the form

$$A\,x \; + \; B\,y \; + \; C\,z \quad = \quad D$$

where $(A,\, B,\, C) \neq (0,0,0)$, then the natural choice for the Gauss map is the normalized gradient

$$\mathbf{N}(\mathbf{p}) \quad = \quad \frac{1}{\sqrt{A^2 + B^2 + C^2}} \cdot (A,\, B,\, C)$$

and accordingly the Gauss map is constant in this case. On the other hand, if one takes the sphere defined by the equation $x^2+y^2+z^2-1 = 0$, then the normalized gradient of the function at (x, y, z) is simply (x, y, z). In this case the image of the Gauss map is the entire sphere. Frequently the image of the Gauss map is somewhere between these two extremes. For example, if we consider the circular cylinder defined by the equation $x^2 + y^2 - 1 = 0$ then the normalized gradient at an arbitrary point (x, y, z) of the cylinder is $(x, y, 0)$, and the image of the Gauss map is the circle in the xy-plane defined by the equation $x^2 + y^2 - 1 = 0$. As a final example consider the parabolic cylinder $y - x^2 = 0$. The gradient at a typical point of this surface has the form $(-2x,\, 1, 0)$, and the corresponding set of unit vectors consists of all points on the unit semicircle in the xy-plane defined by the equation $x^2 + y^2 - 1 = 0$ and the inequality $y > 0$.

IV.2 : The Second Fundamental Form

(Lipschutz, Chapter 9)

The First Fundamental Form carries an enormous amount of information about the geometry of a surface. However, it does not completely characterize surfaces up to rigid motions in the sense of Section I.5 (*i.e*, the existence of a map Φ as in that section such that one of the surfaces is locally the image of the other under Φ). To see this, consider the plane defined by the equation $x = 1$ and the cylinder defined by $x^2 + y^2 = 1$ at the point $(1, 0, 0)$. Clearly these surfaces are not equivalent under a rigid motion. However, if one takes parametrizations near $(1, 0, 0)$ for surfaces of the forms $\mathbf{A}(u, v) = (1, u, v)$ and $\mathbf{B}(u, v) = (\cos u, \sin u, v)$ one obtains the same First Fundamental Form in terms of u and v; in each case one has $\mathbf{I} = du\,du + dv\,dv$. Physically, this reflects the fact that we can roll a flat piece of paper onto a portion of a cylinder without stretching or tearing it. Thus it is clear that we need additional data in order to characterize surfaces locally up to rigid motions. The objective of this unit is to investigate the classical description of this additional information using the Second Fundamental Form. As the name indicates, this is similar to the First Fundamental Form in some key respects, but as one might expect from the organization of these notes, its formulation requires an orientation for the surface and its definition involves the Gauss map.

Suppose that $(\Sigma, \mathbf{N})$ is an oriented surface in $\mathbb{R}^3$. We would like to define a derivative for this map $D\,\mathbf{N}$ such that for each $\mathbf{p} \in \Sigma$ we have a linear transformation $D\,\mathbf{N}(\mathbf{p})$ on the tangent space $T_{\mathbf{p}}(\Sigma)$ and in an appropriate sense $D\,\mathbf{N}(\mathbf{p})$ is a smooth function of $\mathbf{p}$. Formally, one can achieve many of these goals by taking a smooth extension $\mathbf{N}^{\#}$ of $\mathbf{N}$ on some open disk U containing $\mathbf{p}$; given a vector $\mathbf{v}$ in $T_{\mathbf{p}}(\Sigma)$ we can then provisonally define

$$[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v}) \quad = \quad \left[D\,\mathbf{N}^{\#}(\mathbf{p})\right](\mathbf{v})\ .$$

The smoothness of this map is immediate, but it is necessary to check that the right hand side does not depend upon the choice of extension $\mathbf{N}^{\#}$ and that it sends tangent vectors at $\mathbf{p}$ to tangent vectors at $\mathbf{p}$. We shall verify these in order.

LEMMA 1. *The right hand side of the defining equation for* $[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})$ *does not depend upon the choice of extension* $\mathbf{N}^{\#}$.

Proof. Let $\mathbf{y}$ be a regular smooth curve in Σ such that $\mathbf{y}(0) = \mathbf{p}$ and $\mathbf{y}'(0) = \mathbf{v}$. Then the Chain Rule shows that the curve

$$\mathbf{z}(t) \quad = \quad \mathbf{N}\circ\mathbf{y}(t) \quad = \quad \mathbf{N}^{\#}\circ\mathbf{y}(t)$$

is also a smooth curve and $[\mathbf{N}\circ\mathbf{y}]'(0)$ is equal to the right hand side of the defining equation for $[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})$. Since $\mathbf{z}'(0) = [\mathbf{N}\circ\mathbf{y}]'(0)$ does not depend upon the choice of extension, this proves that the definition for $[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})$ also does not depend upon the choice of extension.■

LEMMA 2. *If* $\mathbf{v} \in T_{\mathbf{p}}(\Sigma)$ *then* $[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})$ *also lies in* $T_{\mathbf{p}}(\Sigma)$.

Proof. We shall use the notation of the preceding lemma, so that

$$[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v}) \quad = \quad \mathbf{z}'(0)$$

where $\mathbf{z} = \mathbf{N}\circ\mathbf{y}$. It suffices to show that $\mathbf{z}'(0)$ is perpendicular to $\mathbf{N}(\mathbf{p})$. For each $\mathbf{q} \in \Sigma$ the normal vector $\mathbf{N}(\mathbf{q})$ has unit length by construction, and therefore we know that $|\mathbf{z}(t)|^2 \equiv 1$. If we differentiate this and apply the Leibniz Rule for dot products of vector valued functions, we see that

$$0 \quad = \quad \frac{d}{dt}\,|\mathbf{z}(t)|^2 \quad = \quad 2\,\langle \mathbf{z}(t),\, \mathbf{z}'(t)\,\rangle$$

and if we evaluate the inner product on the right hand side at $t = 0$ we see that it is equal to

$$\langle \mathbf{N}(\mathbf{p}),\, [D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})\,\rangle$$

and therefore we conclude that this inner product vanishes; *i.e.*, the vector $[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})$ is perpendicular to $\mathbf{N}(\mathbf{p})$.■

If we are given a local parametrization of Σ near $\mathbf{p}$ by some regular smooth parametrization $\mathbf{X}$, defined on some connected domain U, then we can use the local formula for $\mathbf{N}$ to describe the map

$$\left[D\left(\mathbf{N}\circ\mathbf{X}\right)\right]\,(u,v)$$

as a linear transformation on $\mathbb{R}^2$ using the identity

$$\mathbf{N}\circ\mathbf{X} \quad = \quad \frac{1}{(\text{length})}\cdot\left(\frac{\partial\mathbf{X}}{\partial u}\right)\;\times\;\left(\frac{\partial\mathbf{X}}{\partial v}\right)$$

but we shall find a better way to compute this map in terms of local coordinates.

Standard terminology. In textbooks and elsewhere the map $-D\,\mathbf{N}(\mathbf{p})$ (note the sign!!) is often called the *Weingarten map* or, particularly as in O'NEILL, the SHAPE OPERATOR for the oriented surface $(\Sigma, \mathbf{N})$. The reason for the term in O'NEILL is related to the Fundamental Theorem in Local Surface Theory, which is discussed in Section V.2 of the notes. The appendix to this section studies this Shape Operator in further detail.

Important special cases. At each point $\mathbf{p}$ the space of tangent vectors at $\mathbf{p}$ has a basis given by the partial derivative vectors $\mathbf{X}_u$ and $\mathbf{X}_v$. By construction, the Weingarten map or Shape Operator at $\mathbf{p}$ is the unique linear transformation which sends the tangent vectors $\mathbf{X}_u$ and $\mathbf{X}_v$ at $\mathbf{p}$ to $-\mathbf{N}_u$ and $-\mathbf{N}_v$ respectively. Lemma 2 implies that the latter two vectors also lie in the space of tangent vectors at $\mathbf{p}$. It follows that the Weingarten map or Shape Operator sends a general tangent vector $a\mathbf{X}_u + b\mathbf{X}_v$ to $-a\mathbf{N}_u - b\mathbf{N}_v$.

The Second Fundamental Form is defined in terms of the Weingarten map or Shape Operator as follows:

Definition. Let $\mathbf{T}^{(2)}(\Sigma)$ be the set of all ordered pairs of points $(\mathbf{v}_1,\, \mathbf{v}_2)$ in $\mathbf{T}(\Sigma)\times\mathbf{T}(\Sigma)$ such that $\tau_\Sigma(\mathbf{v}_1) = \tau_\Sigma(\mathbf{v}_2)$, and let $\mathbf{x}_i$ be the second coordinate of $\mathbf{v}_i$. The *Second Fundamental Form* of Σ is the map $\mathbf{II}_\Sigma$ on $\mathbf{T}^{(2)}(\Sigma)$ sending $(\mathbf{v}_1,\, \mathbf{v}_2)$ to

$$-\;\langle [D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{x}_1),\, \mathbf{x}_2\rangle$$

(note the sign!!!) where as usual $\langle \ldots,\, \ldots\rangle$ denotes the usual inner product of two vectors in $\mathbb{R}^3$.

At this point it is helpful to understand what this means for the basic examples we discussed in the previous section.

Examples. 1. If Σ is a plane and $\mathbf{N}$ is some orientation, then $\mathbf{N}$ is parallel to the normal direction for the plane at all points and therefore $\mathbf{N}$ is at least locally constant by the observations in Section IV.1. In fact, one can take the extension $\mathbf{N}$ also to be constant, and it follows that in this case $D\,\mathbf{N}(\mathbf{p}) = 0$ for all points $\mathbf{p}$ on the plane, and therefore the Second Fundamental Form is also zero at every point of the plane.

2. In contrast, if Σ is the sphere and $\mathbf{N}$ is the standard outward pointing orientation then $\mathbf{N}(\mathbf{p}) = \mathbf{p}$ and therefore $D\,\mathbf{N}(\mathbf{p}) = I$ for all points $\mathbf{p}$ on the sphere; in this case the Second Fundamental Form is the negative of the First Fundamental Form. Of course, if we would replace $\mathbf{N}$ by its negative, then the Second and First Fundamental Forms would be equal.

3. Finally, suppose that Σ is the cylinder defined by $x^2 + y^2 = 1$. In this case $\mathbf{N}(x, y, z) = (x, y, 0)$ and the tangent space at (x, y, z) has an orthonormal basis given by $(-y, x, 0)$, which is the tangent vector to the circle $\mathbf{a}(\theta) = (\cos\theta, \sin\theta, z)$ at (x, y, z), and $(0.0.1)$, which is the tangent vector to the vertical line $\mathbf{b}(t) = (x, y, t)$ at (x, y, z). Direct computation shows that these two vectors are eigenvectors for $D\,\mathbf{N}(x, y, z)$ and the associated eigenvalues are 1 and 0 respectively. It follows that the Second Fundamental Form is given by

$$- \,\langle [D\,\mathbf{N}(\mathbf{p})]\,(p\,\mathbf{a}' \;+\; q\,\mathbf{b}'),\, r\,\mathbf{a}' \;+\; s\,\mathbf{b}'\,\rangle$$

and by the observations in the preceding sentences this is equal to $-\,p\,r$. In particular, the Second Fundamental Form vanishes for some but not all pairs of vectors. At the beginning of this section we mentioned that the First Fundamental Forms for the plane and the cylinder were the same, but we have now seen that their Second Fundamental Forms are different.

Computational formulas

We would also like to understand the behavior of the Second Fundamental Form for the hyperboloid of one sheet given by the equation $x^2 + y^2 - z^2 - 1 = 0$, and in order to do this we need to follow through on our earlier comment about developing a way of computing $D\,\mathbf{N}$ and the Second Fundamental Form in a efficiently using a regular smooth parametrization $\mathbf{X}$. The first point is to observe that the Second Fundamental Form is completely determined by its values for pairs of tangent vectors such that each is either $\mathbf{X}_u$ and $\mathbf{X}_v$ (this includes cases where both vectors in the pair are the same and where the two vectors are different). This is true because

$$\mathbf{II}\,(p\,\mathbf{X}_u \;+\; q\,\mathbf{X}_v),\, r\,\mathbf{X}_u \;+\; s\,\mathbf{X}_v\,) \;\;=$$

$$p\,r\,\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_u) \;+\; p\,s\,\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_v) \;+\; q\,r\,\mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_u) \;+\; q\,s\,\mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_v)\;\;.$$

We then have the following important formulas:

BASIC VALUES FOR SECOND FUNDAMENTAL FORMS. *Suppose we are given a regularly parametrized smooth surface Σ with parametrization $\mathbf{X}$, and assume that the normal vector function is given by $\mathbf{N}$. Then the following identities hold for the second fundamental form:*

[1] $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_u) = \langle \mathbf{N},\,\mathbf{X}_{u,u}\,\rangle$

[2] $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_v) = \langle \mathbf{N},\,\mathbf{X}_{u,v}\,\rangle = \langle \mathbf{N},\,\mathbf{X}_{v,u}\,\rangle = \mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_u)$

[3] $\mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_v) = \langle \mathbf{N},\,\mathbf{X}_{v,v}\,\rangle$

Proof. All the derivations are of a similar nature, so we start with the first one. Since $\mathbf{N}$ is perpendicular to the tangent plane at every point we know that

$$\langle \mathbf{N},\,\mathbf{X}_u\,\rangle \;\;=\;\; 0\;.$$

Next, observe that $D\,\mathbf{N}(\mathbf{p})\,\mathbf{X}_u$ is merely the partial derivative $D_u\,(\mathbf{N}\circ\mathbf{X})$, and we shall stretch our conventions to write this as $D_u\,\mathbf{N}$. Taking partial derivatives of the equation above with respect to u and applying the Leibniz Rule for dot products of vector valued functions, we see that

$$0 \;=\; \langle D_u\,\mathbf{N},\,\mathbf{X}_u\,\rangle \;+\; \langle \mathbf{N},\,\mathbf{X}_{u,u}\,\rangle$$

and since the first summand on the right hand side is the negative of $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_u)$ it follows that the latter is equal to $\langle \mathbf{N},\,\mathbf{X}_{u,u}\,\rangle$ as required. The derivation of the third identity is nearly identical, the only difference being that u is replaced by v in each equation.

Similarly, if we take partial derivatives of both sides of the equation

$$\langle \mathbf{N},\,\mathbf{X}_u\,\rangle \;=\; 0$$

with respect to v we conclude that $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_v) = \langle \mathbf{N},\,\mathbf{X}_{u,v}\,\rangle$. Furthermore, if we interchange the roles of u and v in this argument we also see that $\mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_u) = \langle \mathbf{N},\,\mathbf{X}_{v,u}\,\rangle$. Under the assumption that the regular surface parametrization has continuous second partial derivatives, we know that $\mathbf{X}_{u,v} = \mathbf{X}_{v,u}$ and using this we see that $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_v) = \mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_u)$. This completes the derivation.■

Notational conventions. In the literature and textbooks the quantities $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_u\,)$, $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_v\,) = \mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_u\,)$ and $\mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_v\,)$ are often denoted by e, f and g respectively or by L, M and N respectively; in these notes we shall use the first notation in order to avoid confusion between N and $\mathbf{N}$. If one writes a typical tangent vector in the form $\mathbf{X_u}\,du + \mathbf{X_v}\,dv$ where du and dv are viewed as scalars, then this yields the classical expression for the Second Fundamental Form:

$$\mathbf{II}\,(\mathbf{X}_u\,du + \mathbf{X}_v\,dv,\,\mathbf{X}_u\,du + \mathbf{X}_v\,dv\,) \;=$$

$$e(u,v)\,du\,du \;+\; 2\,f(u,v)\,du\,dv \;+\; g(u,v)\,dv\,dv$$

Example. We shall apply all this to describe the Second Fundamental Form for the hyperboloid of one sheet described above. If we let $h(x,y,z) = x^2 + y^2 - z^2 - 1$ then $\nabla h(x,y,z) = 2\,(x,y,-z)$ and therefore the unit normal is given by

$$\mathbf{N}(x,y,z) = \frac{1}{\sqrt{x^2+y^2+z^2}}\cdot(x,y,-z)\;.$$

This can be simplified slightly by noting that $x^2+y^2+z^2 = 1+2\,z^2$ for points on the surface, but it will also be useful for us to let $\Omega(x,y,z) = (x,y,-z)$, so that $\Omega(x,y,z)$ is a positive multiple of the unit normal $\mathbf{N}$ described above.

We shall use the following standard parametrization for the hyperboloid of one sheet:

$$\mathbf{X}(u,v) \;=\; \big(\cos u\,\cosh v,\,\sin u\,\cosh v,\,\sinh v\big)$$

One can then describe the normal vector Ω by the following formula:

$$\Omega(u,v) \;=\; \big(\cos u\,\cosh v,\,\sin u\,\cosh v,\,-\,\sinh v\big)$$

One can compute the second partial derivatives of $\mathbf{X}$ in a very direct manner, and if one takes inner products with Ω one obtains the following results:

$$\langle \Omega, \mathbf{X}_{u,u} \rangle = -\cosh^2 v$$

$$\langle \Omega, \mathbf{X}_{u,v} \rangle = \langle \Omega, \mathbf{X}_{v,u} \rangle = 0$$

$$\langle \Omega, \mathbf{X}_{v,v} \rangle = 1$$

Therefore it follows that the coefficients $e(u, v)$ are always negative, the coefficients $f(u, v)$ are always zero and the coefficients $g(u, v)$ are always positive. Thus the Second Fundamental Form in this case looks quite different from those in the cases previously described.

Since many surfaces are expressed as graphs of functions of two variables, we shall describe the First and Second Fundamental Forms for a surface defined as the graph of a smooth function $z = h(x, y)$. In order to state these formulas concisely we shall let $\alpha(x, y) = \sqrt{1 + h_x^2 + h_y^2}$.

$$E = 1 + h_x^2$$

$$F = h_x \, h_y$$

$$G = 1 + h_y^2$$

$$e = h_{x,x}/\alpha$$

$$f = h_{x,y}/\alpha$$

$$g = h_{y,y}/\alpha$$

All of these formulas may be verified directly using the identities established above and the fact that the regular smooth parametrization of the surface is given by $\mathbf{X}(x, y) = (x, y, \, h(x, y)\,)$.

Appendix: More on shape Operators

Since the Shape Operator plays such a central role in O'Neill and similar modern approaches to the subject, it seems extremely worthwhile to include a few more identities involving it that are useful for computational purposes. We shall also state explicitly the symmetry property of the Second Fundamental Form which plays a central role in Sections IV.3 – IV.5 (and also in Unit V).

Formal definition of the Shape Operator

Since the Shape Operator was only described informally in the preceding discussion, we shall begin with a more systematic approach. Since we are interested in local formulas here, we shall assume that our surface Σ is the image of some smooth parametrization $\mathbf{X}$ which is defined on a connected domain and is 1–1. Suppose that $\mathbf{N}^{\Sigma}$ is an orientation for Σ. If we take the standard orientation associated to the parametrization given by $\mathbf{N}^{\mathrm{local}}$ which is the unit vector pointing in the same direction as

$$\frac{\partial \mathbf{X}}{\partial u} \times \frac{\partial \mathbf{X}}{\partial v}$$

Then for each point $\mathbf{p} = \mathbf{X}(u, v)$ on Σ we have

$$\mathbf{N}^{\Sigma}\big(\,\mathbf{X}(u, v)\,\big) \quad = \quad \varepsilon(\mathbf{p}) \cdot \mathbf{N}^{\mathrm{local}}\,(u, v)$$

where $\varepsilon = \pm 1$ and is continuous. As noted before, it follows that ε is locally constant, so at least if we cut down the domain of $\mathbf{X}$ to a small open disk containing some point (u_0, v_0) we can assume that ε is constant, and in this case for the sake of convenience we shall assume that $\varepsilon = 1$.

The definition of the Shape Operator involves the notion of tangent space to Σ defined at the beginning of Section III.4; by construction an element of the tangent space $T(\Sigma)$ is a pair $(\mathbf{p}, \mathbf{q})$ consisting of a point $\mathbf{p} \in \Sigma$ and a tangent vector $\mathbf{q}$ to Σ at $\mathbf{p}$ (in other words, there is a smooth curve γ in $\mathbf{R}^3$ such that $\gamma(0) = \mathbf{p}$, the image of γ is contained in Σ, and $\gamma'(0) = \mathbf{q}$). We would like to define a mapping W_0 from $T(\Sigma)$ to $\mathbf{R}^3$ with the following properties:

(1) In the setting above, W maps $(\mathbf{p}, \mathbf{q}) \in T(\Sigma)$ to

$$- \left(\mathbf{N}^{\Sigma} \circ \gamma \right)' (0)$$

where γ is given as above.

(2) The function W is linear in the second variable $\mathbf{q}$ when $\mathbf{p}$ is held constant.

(3) For each $(\mathbf{p}, \mathbf{q}) \in T(\Sigma)$ the vector $W(\mathbf{p}, \mathbf{q})$ lies in the space of tangent vectors $T_{\mathbf{p}}(\Sigma)$ to Σ at $\mathbf{p}$.

(4) The mapping W has reasonable continuity and differentiability properties.

If we have such a map W, then the SHAPE OPERATOR is formally defined by the formula

$$\mathbf{S}(\mathbf{p}, \mathbf{q}) \quad = \quad (\, \mathbf{p}, \, W(\mathbf{p}, \mathbf{q})\,)$$

and for each $\mathbf{p}$ the associated linear map from $T_{\mathbf{p}}(\Sigma)$ to itself will be denoted by $\mathbf{S}_{\mathbf{p}}$.

The third property is shown explicitly in the notes, and we shall consider the remaining ones in order.

Independence of choice of curve. In order to show this, we shall use the parametrization. Let $\mathbf{p} = \mathbf{X}(u, v)$; the Normal Thickening Principle in Section III.2 implies that, at least locally, we can write $\gamma = \mathbf{X} \circ \alpha$ for some smooth curve α in the domain of $\mathbf{X}$ such that $\alpha(0) = (u, v)$ and $\mathbf{w} = \alpha'(0)$ satisfies $\mathbf{q} = D\mathbf{X}(u, v)[\mathbf{w}]$. It then follows from the defining formulas and the Chain Rule that

$$\left(\mathbf{N}^{\Sigma} \circ \gamma\right)' (0) \quad = \quad \left(\mathbf{N}^{\Sigma} \circ \mathbf{X} \circ \alpha\right)' (0) \quad =$$

$$\left(\mathbf{N}^{\mathrm{local}} \circ \alpha\right)' (0) \quad = \quad D\mathbf{N}^{\mathrm{local}}(u, v) \alpha'(0) \quad = \quad D\mathbf{N}^{\mathrm{local}}(u, v) \mathbf{w} \ .$$

If we have chosen some other curve γ_1 with the right properties, then we would obtain a similar curve α_1 in the domain of $\mathbf{X}$ such that $\alpha'(0) = \alpha_1'(0) = \mathbf{w}$. Therefore it follows that we would obtain the same tangent vector if we used γ_1 instead of γ, and this proves (1).■

Linearity in the second variable. This is an immediate consequence of the equations displayed above. If we have curves γ_1 and γ_2 in the surface through $\mathbf{p}$, then we have the corresponding curves α_1 and α_2 in the domain of $\mathbf{X}$ through (u, v). Consider the curve

$$\alpha_0(t) \quad = \quad \alpha_1(t) \; + \; \alpha_2(t) \; - \; (u, v) \ .$$

Then we have $\alpha_0(0) = (u, v)$ so that the curve lies in the domain of $\mathbf{X}$ for t sufficiently close to 0, and in addition we have the identity

$$\alpha_0'(0) \quad = \quad \alpha_1'(0) \; + \; \alpha_2'(0) \ .$$

If we let $\gamma_0 = \mathbf{X} \circ \alpha_0$, then the Chain rule implies

$$\gamma_0'(0) \quad = \quad \gamma_1'(0) \; + \; \gamma_2'(0)$$

and the additivity property of W follows immediately from this and the previously displayed formulas. Similarly, if we are given γ and α as before and c is a scalar, then the curve $\beta(t) = \gamma(ct)$ satisfies $\beta'(0) = c\,\gamma'(0)$ and $\beta(t) = \mathbf{X}\circ\alpha(ct)$, which implies the homogeneity of W with respect to scalar multiplication. These observations show that W is linear in $\mathbf{q}$ if $\mathbf{p}$ is held fixed.∎

Continuity and smoothness properties. We shall be somewhat sketchy about these because although they may be intuitively clear, writing out all the details is lengthy and not particularly instructive; furthermore, we can often avoid using these properties directly in elementary work. Since continuity and smoothness only depend on the behavior of a function very close to an arbitrary point, we shall focus on an arbitrary point $\mathbf{p}_0 = \mathbf{X}(u_0, v_0)$ of the surface and all point sufficiently close to $\mathbf{p}_0$ so that an smooth inverse to the normal thickening map Φ (from Section III.2) can be defined. We then have a nonzero vector valued function $\mathbf{G}_0 = \Phi_1 \times \Phi_2$, where Φ_j is the j^{th} partial derivative vector, and we take $\mathbf{G}$ to be the unit vector pointing in the same direction as $\mathbf{G}_0$. It follows that

$$W(\mathbf{p},\mathbf{q}) \quad = \quad -D\mathbf{N}^{\text{local}}\big(\,\Phi^{-1}(\mathbf{p})\,\big)\mathbf{z}$$

where

$$\mathbf{z} \quad = \quad D\Phi^{-1}(\mathbf{p})\mathbf{q}$$

and this description yields all the continuity and smoothness properties one could hope for.∎

Local formula for the shape operator

The derivation of the first property yields the following description of the shape operator in terms of the parametrization:

LOCAL FORMULA. *If $\mathbf{p} = \mathbf{X}(u,v)$ and $\mathbf{q} = D\mathbf{X}(u,v)\mathbf{w}$ in the setting above, then*

$$\mathbf{S}(\mathbf{p},\mathbf{q}) \quad = \quad \big(\,\mathbf{X}(u,v),\, -D\mathbf{N}^{\text{local}}(u,v)\mathbf{w}\,\big)\ .$$

This is merely a reformulation of the previously displayed identity.∎

Computing the Second Fundamental Form

Once again, the global object on Σ and the local object defined on the domain of a parametrization are often identified with each other during informal discussions, so we shall begin by describing and comparing the local and global versions.

GLOBAL VERSION. Given two points $(\mathbf{p},\mathbf{a})$ and $(\mathbf{p},\mathbf{b})$ in $T(\Sigma)$ representing tangent vectors to the same point, the global Second Fundamental Form is defined by

$$\mathbf{II}^{\Sigma}\big(\,(\mathbf{p},\mathbf{a}),\,(\mathbf{p},\mathbf{b})\,\big) \quad = \quad \langle\, W(\mathbf{p},\mathbf{a}),\, \mathbf{b}\,\rangle$$

where $\langle -, - \rangle$ denotes the usual inner product in $\mathbf{R}^3$ and the Second Fundamental Form is often written more compactly as as $\mathbf{II}^{\Sigma}_{\mathbf{p}}(\mathbf{a},\mathbf{b})$.

LOCAL VERSION. This is defined for all (u,v) in the domain of $\mathbf{X}$ and all vectors $\mathbf{y}$ and $\mathbf{z}$ in $\mathbf{R}^2$ by the classical formula

$$\mathbf{II}^{\text{local}}_{(u,v)}(\mathbf{y},\mathbf{z}) \quad = \quad -\langle\, D\mathbf{N}^{\text{local}}(u,v)\mathbf{y}, D\mathbf{X}(u,v)\mathbf{z}\,\rangle$$

where $\langle -, - \rangle$ denotes the usual inner product in $\mathbf{R}^2$. The corresponding classical formula for the First Fundamental Form is

$$\mathbf{I}^{\text{local}}_{(u,v)}(\mathbf{y},\mathbf{z}) \quad = \quad -\langle\, D\mathbf{X}(u,v)\mathbf{y}, D\mathbf{X}(u,v)\mathbf{z}\,\rangle\ .$$

The previous observations yield the following identity for passing from one version of the Second Fundamental Form to the other:

COMPATIBILITY RELATION. *In the preceding discussion, suppose that* $\mathbf{p} = \mathbf{X}(u,v)$*,* $D\mathbf{X}(u,v)\mathbf{y} = \mathbf{a}$ *and* $D\mathbf{X}(u,v)\mathbf{z} = \mathbf{b}$*. Then we have*

$$\mathbf{II}^{\mathrm{local}}_{(u,v)}(\mathbf{y},\mathbf{z}) \quad = \quad \mathbf{II}^{\Sigma}_{\mathbf{p}}(\mathbf{a},\mathbf{b}) \ . \blacksquare$$

Symmetry property of the Second Fundamental Form

One advantage of the local version of the Second Fundamental Form is that it quickly yields the following basic symmetry property.

PROPOSITION. *In the setting above we have*

$$\mathbf{II}^{\mathrm{local}}_{(u,v)}(\mathbf{y},\mathbf{z}) \quad = \quad \mathbf{II}^{\mathrm{local}}_{(u,v)}(\mathbf{z},\mathbf{y})$$

for all u*,* v*,* $\mathbf{y}$ *and* $\mathbf{z}$*.*

COROLLARY. *In the setting above we have*

$$\mathbf{II}^{\Sigma}_{\mathbf{p}}(\mathbf{a},\mathbf{b}) \quad = \quad \mathbf{II}^{\Sigma}_{\mathbf{p}}(\mathbf{b},\mathbf{a})$$

for all $\mathbf{p}$*,* $\mathbf{a}$ *and* $\mathbf{b}$*.*

The corollary follows from the proposition and the compatibility relation between the local and global versions of the Second Fundamental Form.

Proof of Proposition. As in the result at the beginning of Section IV.3, it suffices to prove this when $\mathbf{y}$ and $\mathbf{z}$ are the standard unit vectors $\mathbf{e}_1$ and $\mathbf{e}_2$. It will be convenient for us to denote the partial derivatives of $\mathbf{N} = \mathbf{N}^{\mathrm{local}}$ and $\mathbf{X}$ by $\mathbf{N}_j$ and $\mathbf{X}_j$ here.

By definition we have

$$\mathbf{II}^{\mathrm{local}}_{(u,v)}(\mathbf{e}_1,\mathbf{e}_2) \quad = \quad -\langle\, \mathbf{N}_1,\, \mathbf{X}_2\,\rangle$$

and by the computations of Section IV.2 we know that the right hand side is equal to $\langle\, \mathbf{N},\, \mathbf{X}_{2,1}\,\rangle$. Similarly, we have

$$\mathbf{II}^{\mathrm{local}}_{(u,v)}(\mathbf{e}_2,\mathbf{e}_1) \quad = \quad -\langle\, \mathbf{N}_2,\, \mathbf{X}_1\,\rangle$$

and by the computations of Section IV.2 we know that the right hand side is equal to $\langle\, \mathbf{N},\, \mathbf{X}_{1,2}\,\rangle$. Since $\mathbf{X}_{2,1} = \mathbf{X}_{1,2}$ by equality of mixed partial derivatives, it follows that we have proven the symmetry condition in the special case, and as noted before the general case follows from this.$\blacksquare$

Since the Second Fundamental Form is defined in terms of the standard inner product on $\mathbf{R}^2$ and the shape operator, we also have the following consequence.

PROPOSITION. *In the setting above, for each* $\mathbf{p} \in \Sigma$ *the linear transformation* $\mathbf{S}_{\mathbf{p}}$ *defined on* $T_{\mathbf{p}}(\Sigma)$ *by the shape operator has the following* SELF – ADJOINTNESS *property:*

$$\langle\, \mathbf{S}_{\mathbf{p}}(\mathbf{a}), \mathbf{b}\,\rangle \quad = \quad \langle\, \mathbf{S}_{\mathbf{p}}(\mathbf{b}), \mathbf{a}\,\rangle$$

Proof. This follows from the previous results because the left and right hand sides are equal to the values of the Second Fundamental Forms at $(\mathbf{a},\mathbf{b})$ and $(\mathbf{b},\mathbf{a})$ respectively.$\blacksquare$

The self-adjointness identity is extremely important, and it is used extensively in the remainder of the notes. In the next section we shall develop some algebraic tools that will allow us to use self-adjointness effectively.

IV.3 : Quadratic forms and adjoint transformations

(No suitable text reference)

The First and Second Fundamental Forms are examples of quadratic forms on a real vector space with an inner product. It is particularly useful to study some aspects of the Second Fundamental Form using a few basic algebraic facts about such quadratic forms, so we shall summarize what is needed here. For our purposes it will suffice to restrict our attention to quadratic forms on 2-dimensional real inner product spaces.

The following result is an easy algebraic exercise:

PROPOSITION. *Let V be a 2-dimensional real inner product space with basis vectors $\mathbf{x}$ and $\mathbf{y}$, and let $T : V \to V$ be a linear transformation from V to itself. If $\langle T(\mathbf{x}), \mathbf{y}\rangle = \langle \mathbf{x}, T(\mathbf{y})\,\rangle$, then $\langle T(\mathbf{w}), \mathbf{z}\rangle = \langle \mathbf{w}, T(\mathbf{z})\,\rangle$ for all $\mathbf{w}, \mathbf{z} \in V$.*

Proof. Express $\mathbf{w}$ and $\mathbf{z}$ as linear combinations of $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{w} \;=\; p\,\mathbf{x} + q\,\mathbf{y} \qquad \mathbf{z} \;=\; r\,\mathbf{x} + s\,\mathbf{y}$$

We then have

$$\langle T(\mathbf{w}), \mathbf{z}\rangle \;=\; p\,r\,\langle T(\mathbf{x}), \mathbf{x}\rangle \;+\; q\,r\,\langle T(\mathbf{y}), \mathbf{x}\rangle \;+\; p\,s\,\langle T(\mathbf{x}), \mathbf{y}\rangle \;+\; q\,s\,\langle T(\mathbf{w}), \mathbf{z}\rangle$$

and similarly we have

$$\langle \mathbf{w}, T(\mathbf{z})\,\rangle \;=\; p\,r\langle \mathbf{x}, T(\mathbf{x})\,\rangle \;=\; q\,r\,\langle \mathbf{x}, T(\mathbf{y})\,\rangle \;=\; p\,s\,\langle \mathbf{y}, T(\mathbf{x})\,\rangle \;=\; q\,s\,\langle \mathbf{y}, T(\mathbf{y})\,\rangle\ .$$

We always have

$$\langle T(\mathbf{x}), \mathbf{x}\rangle = \langle \mathbf{x}, T(\mathbf{x})\,\rangle$$

and similarly if $\mathbf{y}$ replaces $\mathbf{x}$, so the hypothesis $\langle T(\mathbf{x}), \mathbf{y}\rangle = \langle \mathbf{x}, T(\mathbf{y})\,\rangle$ combines with these to show that $\langle T(\mathbf{w}), \mathbf{z}\rangle = \langle \mathbf{w}, T(\mathbf{z})\,\rangle$ for all $\mathbf{w}, \mathbf{z} \in V$.$\blacksquare$

Linear transformations satisfying the conclusion of the preceding result are said to be *self-adjoint*. If we are given an orthonormal basis $\mathbf{u}$ and $\mathbf{v}$ for our inner product space V and we construct the 2×2 matrix representing T with this orthonormal basis

$$T(\mathbf{u}) \;=\; a\,\mathbf{u} + b\,\mathbf{v} \qquad T(\mathbf{v}) \;=\; c\,\mathbf{u} + d\,\mathbf{v}$$

then T is self-adjoint if and only if

$$c = \langle \mathbf{u}, T(\mathbf{v})\,\rangle \;=\; \langle T(\mathbf{u}), \mathbf{v}\,\rangle \;=\; b$$

or in other words the matrix representing T is *symmetric*. (SKETCH OF PROOF: If T is self-adjoint, then the displayed formula is merely a special case of the general definition for a self-adjoint linear transformation. Conversely, if $\mathbf{w}$ and $\mathbf{z}$ are arbitrary vectors, then we may write them as linear combinations of $\mathbf{u}$ and $\mathbf{v}$, and by combining the displayed formulas with some elementary but slightly messy algebraic expansions we can check directly that $\langle T(\mathbf{w}), \mathbf{z}\rangle = \langle \mathbf{w}, T(\mathbf{z})\,\rangle$.$\blacksquare$).

Every real symmetric matrix has an orthonormal basis of eigenvectors; this is a standard result on matrices, and in the 2×2 case one can see this very easily by computing the characteristic polynomial and noting that it has real roots. Other basic results in linear algebra imply a corresonding result of this sort for self-adjoint linear transformations.

Let V and T be as above. In the next section we shall be interested in finding the maximum and minimum values of the quotient

$$k(\mathbf{x}) \quad = \quad \frac{\langle T(\mathbf{x}),\, \mathbf{x}\,\rangle}{\langle \mathbf{x},\, \mathbf{x}\,\rangle}$$

where $\mathbf{x}$ ranges over all nonzero vectors in V. Let $\mathbf{u}$ and $\mathbf{v}$ be an orthonormal basis of eigenvectors for T, and let α and β be the eigenvalues associated to $\mathbf{u}$ and $\mathbf{v}$ respectively. One of these eigenvalues is greater than or equal to the other, and we shall assume that we have labeled everything so that $\alpha \leq \beta$.

RAYLEIGH'S PRINCIPLE. *The maximum and minimum values of the above expression are the eigenvalues β and α, and these values are attained at the eigenvectors $\mathbf{v}$ and $\mathbf{u}$ respectively.*

Proof. If $\mathbf{x}$ is an arbitrary nonzero vector in V we may write

$$\mathbf{x} \quad = \quad r\,\cos\theta\,\mathbf{u} \;+\; r\,\sin\theta\,\mathbf{v}$$

for some $r > 0$ and $\theta \in \mathbb{R}$. It follows that

$$k(\mathbf{x}) = \alpha\,\cos^2\theta + \beta\,\sin^2\theta\ .$$

Since $\alpha \leq \beta$ it follows that

$$\alpha \quad = \quad \alpha\,\cos^2\theta + \alpha\,\sin^2\theta \quad \leq \quad \alpha\,\cos^2\theta + \beta\,\sin^2\theta \quad \leq \beta\,\cos^2\theta + \beta\,\sin^2\theta \quad = \quad \beta$$

and it also follows that $k(\mathbf{u}) = \alpha$ while $k(\mathbf{v}) = \beta$.∎

Trace and determinant formulas

Recall that the **trace** of a square matrix is equal to the sum of its diagonal entries, and if a matrix is diagonalizable then the trace is equal to the weighted sum of the eigenvalues

$$\sum_{\lambda}\; n(\lambda)\,\lambda$$

(one way to see this is by means of the characteristic polynomial — both numbers are $(-1)^{n-1}$ times the coefficient of t^{n-1} if the matrix in question is $n \times n$). One can then define the trace of a diagonalizable linear transformation on a finite-dimensional vector space by means of the corresponding weighted sum of eigenvalues. This is entirely analogous to the situation for the determinant. For diagonalizable matrices the latter is equal to the weighted product of eigenvalues

$$\prod_{\lambda}\; n(\lambda)\,\lambda$$

and one can define the determinant of a diagonalizable linear transformation on a finite dimensional vector space using this formula.

We shall need information about the following purely algebraic question:

Problem. Suppose that we are given a 2-dimensional real inner product space V, a basis $\mathbf{z}_1$ and $\mathbf{z}_2$ for V, and a self-adjoint linear transformation $T : V \to V$. Suppose that

$$A \quad = \quad \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

is the matrix whose entries are defined by the formulas

$$T(\mathbf{z}_1) \quad = \quad a\,\mathbf{x}_1 \; + \; b\,\mathbf{z}_2 \qquad \text{and} \qquad T(\mathbf{z}_1) \quad = \quad a\,\mathbf{x}_1 \; + \; b\,\mathbf{z}_2 \; .$$

Express the entries of A in terms of the inner products $\langle \mathbf{z}_i, \, \mathbf{z}_j \rangle$ and $\langle T(\mathbf{z}_i), \, \mathbf{z}_j \rangle$ where $1 \le i, j \le 2$.

Motivated by our terminology for the First and Second Fundamental Forms, we shall denote the various inner products as indicated in the matrices below:

$$\begin{pmatrix} \langle \mathbf{z}_1, \, \mathbf{z}_1 \rangle & \langle \mathbf{z}_1, \, \mathbf{z}_2 \rangle \\ \langle \mathbf{z}_2, \, \mathbf{z}_1 \rangle & \langle \mathbf{z}_2, \, \mathbf{z}_2 \rangle \end{pmatrix} \quad = \quad \begin{pmatrix} E & F \\ F & G \end{pmatrix} \qquad\qquad \begin{pmatrix} \langle T(\mathbf{z}_1), \, \mathbf{z}_1 \rangle & \langle T(\mathbf{z}_1), \, \mathbf{z}_2 \rangle \\ \langle T(\mathbf{z}_2), \, \mathbf{z}_1 \rangle & \langle T(\mathbf{z}_2), \, \mathbf{z}_2 \rangle \end{pmatrix} \quad = \quad \begin{pmatrix} e & f \\ f & g \end{pmatrix}$$

Direct calculation then yields the following equations:

$$\begin{aligned}
e \quad &= \quad \langle T(\mathbf{z}_1), \, \mathbf{z}_1 \rangle \quad = \quad \langle T(a\,\mathbf{z}_1 \, + \, b\,\mathbf{z}_2), \, \mathbf{z}_1 \rangle \quad = \quad a\,E \, + \, b\,F \\
f \quad &= \quad \langle T(\mathbf{z}_2), \, \mathbf{z}_1 \rangle \quad = \quad \langle T(c\,\mathbf{z}_1 \, + \, d\,\mathbf{z}_2), \, \mathbf{z}_1 \rangle \quad = \quad c\,E \, + \, d\,F \\
f \quad &= \quad \langle T(\mathbf{z}_1), \, \mathbf{z}_2 \rangle \quad = \quad \langle T(a\,\mathbf{z}_1 \, + \, b\,\mathbf{z}_2), \, \mathbf{z}_2 \rangle \quad = \quad a\,F \, + \, b\,G \\
g \quad &= \quad \langle T(\mathbf{z}_2), \, \mathbf{z}_2 \rangle \quad = \quad \langle T(c\,\mathbf{z}_1 \, + \, d\,\mathbf{z}_2), \, \mathbf{z}_1 \rangle \quad = \quad c\,F \, + \, d\,G
\end{aligned}$$

These equations are equivalent to the following matrix equation:

$$\begin{pmatrix} e & f \\ f & g \end{pmatrix} \quad = \quad \begin{pmatrix} a & c \\ b & d \end{pmatrix} \cdot \begin{pmatrix} E & F \\ F & G \end{pmatrix}$$

If we now assume that the matrix with entries E, F and G is invertible (as it is in the case of the First Fundamental Form), then one can solve for A and obtain descriptions of its entries in terms of the entries of the other two matrices. These lead directly to the identities we want:

TRACE AND DETERMINANT FORMULAS. *The determinant of A is equal to*

$$\frac{e\,g - f^2}{E\,G - F^2}$$

and the trace of A is equal to

$$\frac{e\,G - 2\,f\,F + g\,E}{E\,G - F^2} \; .$$

Derivation. The first of these follows from the matrix equation and the fact that $\det(B_1 B_2) = \det B_1 \cdot \det B_2$. For the second formula we need to compute

$$\begin{pmatrix} e & f \\ f & g \end{pmatrix} \cdot \begin{pmatrix} E & F \\ F & G \end{pmatrix}^{-1}$$

and take the sum of its diagonal entries. By Cramer's Rule the inverse is given by

$$\frac{1}{EG - F^2} \cdot \begin{pmatrix} G & -F \\ -F & E \end{pmatrix}$$

and if substitute this into the preceding formula, compute the product, and add the diagonal entries then we obtain the expression for the trace in the formula.■

Quadratic forms in three or more variables

Basic results of linear algebra state that quadratic forms with real coefficients can always be diagonalized as in the 2×2 case. A detailed account of these results appears in Sections IV.3 and V.1 of the following online notes:

`http://math.ucr.edu/∼res/math132/linalgnotes.pdf`

Section V.2 of the latter also describes a fundamental geometric application of this result to conic sections and quadric surfaces; namely, every "genuine" conic is congruent to one of the examples in the first list below, and every "genuine" quadric surface is congruent to one in the second list:

NONDEGENERATE CONICS:

1. *Ellipses (and circles)* of the form $a^2x^2 + b^2y^2 = 1$, where $a, b > 0$.

2. *Hyperbolas* of the form $a^2x^2 + b^2y^2 = 1$, where $a, b > 0$.

3. *Parabolas* of the form $a^2x^2 = by$, where $a > 0$ and $b \neq 0$.

NONDEGENERATE QUADRIC SURFACES:

1. *Ellipsoids (and spheres)* of the form $a^2x^2 + b^2y^2 + c^2z^2 = 1$, where $a, b, c > 0$.

2. *One sheeted hyperboloids* of the form $a^2x^2 + b^2y^2 - c^2z^2 = 1$, where $a, b, c > 0$.

3. *Elliptic (and circular) cones* of the form $a^2x^2 + b^2y^2 + c^2z^2 = 0$, where $a, b, c > 0$.

4. *Two sheeted hyperboloids* of the form $a^2x^2 - b^2y^2 - c^2z^2 = 1$, where $a, b, c > 0$.

5. *Elliptic paraboloids* of the form $a^2x^2 + b^2y^2 = cz$, where $a, b > 0$ and $c \neq 0$.

6. *Elliptic (and circular) cylinders* of the form $a^2x^2 + b^2y^2 = 1$, where $a, b > 0$.

7. *Hyperbolic paraboloids* of the form $a^2x^2 - b^2y^2 = 1$, where $a, b > 0$.

8. *Hyperbolic cylinders* of the form $a^2x^2 + b^2y^2 = 1$, where $a, b > 0$.

9. *Parabolic cylinders* of the form $a^2x^2 = by$, where $a > 0$ and $b \neq 0$.

The term "genuine" is meant to exclude degenerate or singular possibilities including a pair of lines or planes, one point sets defined by equations like $x^2 + y^2 + z^2 = 0$, and examples involving equations with no real solutions such as $x^2 + y^2 + z^2 + 1 = 0$.

It is possible to prove that the above listings separate conics and quadrics into distinct congruence types; in other words, there is no conic or quadric that is simultaneously congruent to objects of types $\mathbf{m}$ and $\mathbf{n}$ where $\mathbf{m} \neq \mathbf{n}$. However, an efficient proof of this fact requires material beyond the scope of this course. For the sake of completeness, a reference is given below (but the discussion is at a relatively advanced level).

`http://math.ucr.edu/ res/progeom/quadrics1.pdf`

IV.4 : Normal, Gaussian and mean curvature

(Lipschutz, Chapter 9)

One approach to studying the curvature properties of surfaces is to consider the curvature properties curves formed by the intersection of a surface with some plane containing a point on the surface. In particular, if one wants to study the curvature properties of an oriented surface $(\Sigma, \mathbf{N})$ at some point $\mathbf{p} \in \Sigma$, one might consider the curves formed by intersecting Σ with all planes containing the normal line to $\mathbf{p}$ and attempt to describe their curvatures. In fact, this approach leads directly to the basic notions of curvature for oriented surfaces in $\mathbb{R}^3$.

It will be convenient to review some concepts from the first unit of the course. If we are given a regular smooth curve $\mathbf{y}$ in $\mathbb{R}^3$ such that $\mathbf{y}(0) = \mathbf{p}$, then we shall let s denote the modified arc length parametrization such that $s(0) = 0$ and $s'(t) = |\mathbf{y}'(t)|$. Then one has the unit tangent vector function

$$\mathbf{T} \quad = \quad \frac{d\,\mathbf{y}}{d\,s} \quad = \quad \frac{\mathbf{y}'(t)}{s'(t)}$$

and the associated curvature vector function

$$\mathbf{k}(s) \quad = \quad \frac{d\,\mathbf{T}}{d\,s} \quad = \quad \frac{\mathbf{T}\,'(t)}{s'(t)}$$

that is perpendicular to $\mathbf{T}$ and whose magnitude is equal to the curvature of of $\mathbf{y}$ at a given parameter value.

Definition. Let $\mathbf{y}$ be a smooth curve in Σ such that $\mathbf{y}(0) = \mathbf{p}$; if $\mathbf{X}$ is a regular smooth parametrization of Σ at $\mathbf{p}$ then we may write $\mathbf{y}(t) = \mathbf{X}(\,u(t),\, v(t)\,)$ for suitable smooth functions u and v, at least if t is sufficiently close to 0. The *normal curvature vector* $\mathbf{k}_n$ for the curve is then given by

$$\mathbf{k}_n(s) \quad = \quad \big\langle\, \mathbf{k}(s),\, \mathbf{N}(\,\mathbf{y}(s)\,)\,\big\rangle \;\cdot\; \mathbf{N}(\,\mathbf{y}(s)\,)$$

and the *normal curvature* of $\mathbf{y}$ with respect to Σ is given by

$$\kappa_n \quad = \quad \langle \mathbf{k},\, \mathbf{N}\,\rangle\ .$$

The normal curvature vector and the normal curvature are related by the equation $\mathbf{k}_n = \kappa_n \cdot \mathbf{N}$.

IMPORTANT SPECIAL CASE. Suppose that we are given a curve $\mathbf{y}$ defined as the intersection of Σ with the plane through $\mathbf{p}$ that contains the normal line to Σ and the tangent line through $\mathbf{p}$ that is parallel to the nonzero vector $\mathbf{v} \in T_{\mathbf{p}}(\Sigma)$. Then the curvature vector for $\mathbf{y}$ at parameter value is perpendicular to $\mathbf{v}$ and lies in the plane containing this vector and $\mathbf{N}(\mathbf{p})$, and accordingly $\mathbf{k}(0)$ is a scalar multiple of $\mathbf{N}(\mathbf{p})$. In this case *the absolute value of the normal curvature is the ordinary curvature* of $\mathbf{y}$ at parameter value 0.

The following crucial result allows us to describe the normal curvature in relatively familiar terms.

MEUSNIER'S THEOREM. *The normal curvature vector* $\mathbf{k}_n$ *and the normal curvature* κ_n *at* $\mathbf{p}$ *only depend upon* $\mathbf{y}'(0) = \mathbf{w}$*, and in fact we have the formula*

$$\kappa_n \quad = \quad \frac{\mathbf{II}(\mathbf{w},\, \mathbf{w})}{\mathbf{I}(\mathbf{w},\, \mathbf{w})}\ .$$

Of course, this is just the sort of expression that we considered at the end of the previous section.

Historical footnote. Born on June 19, 1754, at Tours, France, JEAN BAPTISTE MARIE MEUSNIER is known for his ideas on designing airships and his career as a military officer as well as his results on the differential geometry of surfaces. He was a student of G. Monge at the École Royale du Génie in Mézières, and he was the first person to envision an elongated airship as an alternative to a spherical balloon. His suggestion of an elliptical-shaped airship was advanced in 1784, just weeks after the first flights of hot air balloons by the Montgolfier brothers. Henri Giffard adopted much of Meusnier's design in his first successful powered airship. Meusnier also played a key role in the organization of the army of the First French Republic; he was severely wounded during a battle between the French and Prussians at Cassel (near Mainz, Germany), and died on June 13, 1793.

Proof of Meusnier's Theorem. We know that the unit tangent vector $\mathbf{T}$ to the curve is perpendicular to the unit normal vector $\mathbf{N}$ to the surface because the curve lies in the surface. Differentiating both sides of the expression $0 = \mathbf{T}\cdot\mathbf{N}$ and applying the Leibniz Rule, we see that

$$0 \;=\; \frac{d\,\mathbf{T}}{d\,t}\cdot\mathbf{N} \;+\; \mathbf{N}\cdot\frac{d\,\mathbf{T}}{d\,t}$$

This leads to the following string of equations:

$$\kappa_n \;=\; \mathbf{k}\cdot\mathbf{N} \;=\; \frac{d\,\mathbf{T}}{ds}\cdot\mathbf{N} \;=\; \frac{1}{s'(t)}\cdot\left(\frac{d\,\mathbf{T}}{d\,t}\cdot\mathbf{N}\right) \;=\; -\,\frac{1}{s'(t)}\cdot\left(\mathbf{T}\cdot\frac{d\,\mathbf{N}}{d\,t}\right) \;=$$

$$-\,\frac{1}{s'(t)^2}\cdot\left(\frac{d\,\mathbf{y}}{d\,t}\cdot\frac{d\,\mathbf{N}}{d\,t}\right) \;=\; -\,\frac{1}{|\mathbf{y}'(t)|^2}\cdot\left(\frac{d\,\mathbf{y}}{d\,t}\cdot\frac{d\,\mathbf{N}}{d\,t}\right)$$

and by the Chain Rule the last expression is equal to

$$-\,\frac{(u'\,\mathbf{X}_u \;+\; v'\,\mathbf{X}_v)\cdot(u'\,\mathbf{N}_u \;+\; v'\,\mathbf{N}_v)}{|u'\,\mathbf{X}_u \;+\; v'\,\mathbf{X}_v|^2}\ .$$

The denominator of this expression is equal to the First Fundamental Form at $(\mathbf{w},\,\mathbf{w})$. Furthermore, the numerator is equal to the inner product of $\mathbf{w}$ and $[D\,\mathbf{N}(\mathbf{p})](\mathbf{w})$, or equivalently the negative of the value of the Second Fundamental Form at $(\mathbf{w},\,\mathbf{w})$. It follows that κ_n is the quotient of the Second Fundamental Form by the First evaluated at $(\mathbf{w},\,\mathbf{w})$.■

By the results of the preceding section, the normal curvatures attain maximum and minimum values, these are realized at the eigenvectors of $-D\,\mathbf{N}(\mathbf{p})$, and the values of the ratio in Meusnier's Theorem are equal to the eigenvalues of $-D\,\mathbf{N}(\mathbf{p})$. The average of these eigenvalues is called the *mean curvature* and the product is called the *Gaussian curvature*. Classically these quantities are denoted by H and K respectively.

LOCAL FORMULAS. *If the oriented surface $(\Sigma,\,\mathbf{N})$ is given by a regular smooth parametrization $\mathbf{X}$, then the mean and Gaussian curvatures H and K are given by the following formulas:*

$$H \;=\; \frac{e\,G - 2\,f\,F + g\,E}{2\,(E\,G - F^2)}\ .$$

$$K = \frac{e\,g - f^2}{E\,G - F^2}$$

Proof. These are immediate consequences of the trace and determinant formulas at the end of the previous section.■

Once again we shall consider our standard examples and describe their mean and Gaussian curvatures. For the plane, we know that the Second Fundamental Form is identically zero, and therefore it follows that both the mean and Gaussian curvatures are zero everywhere. Suppose now that we consider the sphere defined by the equation $x^2 + y^2 + z^2 - r^2 = 0$ where $r > 0$. In this case the unit normal is given by $(x/r,\, y/r,\, z/r)$, so the Second Fundamental Form is just $-1/r$ times the First Fundamental Form. It follows that the normal curvature of every smooth curve through a point on the sphere is equal to $-1/r$, which in turn means that the mean curvature is equal to $-1/(2\,r)$ at each point and the Gaussian curvature is equal to $1/r^2$ at each point. Consider next the cylinder defined by $x^2 + y^2 = 1$. We noted that the eigenvalues of the map $D\,\mathbf{N}$ in this case were equal to 0 and 1, and therefore the mean and Gaussian curvatures in this case are equal to $\frac{1}{2}$ and 0 respectively. Finally, one can use our previous discussion for the hyperbolic paraboloid to find its mean and Gaussian curvatures at each point. One important new feature is that these quantities are no longer constants. We shall not go through all the details of this case but simply note that the mean and Gaussian curvatures for the hyperbolic paraboloid are negative at every point.

Note. The Gaussian curvature can in fact be defined for nonorientable surfaces. This is based upon the following observations:

(*i*) Locally the surface is given by a regular parametrization.

(*ii*) Surfaces given by a single regular parametrization are orientable, and locally they have exactly two orientations, one of which is the negative of the other.

(*iii*) One can use the preceding methods to compute the Gaussian curvature near a point $\mathbf{p}$, and the value is the same for both orientations near $\mathbf{p}$ essentially because the product of the eigenvalues for a diagonalizable 2×2 matrix A is the same as the product for $-A$.

In particular, this means that we may define the Gaussian curvature on the Möbius strip even though there is no globally defined smooth unit normal. These considerations also show that if Σ has an orientation, then the Gaussian curvature does not depend upon the specific choice of orientation.

Interpreting the sign of Gaussian curvature

As a first step to understanding the meaning of curvature for surfaces, it is important to consider the implications for the shape of the surface Σ if the Gaussian curvature is positive, negative or zero at a point. Since Gaussian curvature is continuous, if it is positive or negative at $\mathbf{p}$ then it is also positive or negative at all points close to $\mathbf{p}$, but if the Gaussian curvature is zero at $\mathbf{p}$ then one does not expect to draw any conclusion at all about the nonnegativity or nonpositivity of the Gaussian curvature near $\mathbf{p}$, and in fact we shall give examples to show that the shape of a surface near a point can vary significantly if the Gaussian curvature at $\mathbf{p}$ is zero.

The first step in using Gaussian curvature to obtain a rough idea of the shape near a point $\mathbf{p}$ is to consider the curves formed by intersection the surface with a plane $\mathbf{Q}$ containing the normal line M through $\mathbf{p}$. Let L be the tangent line to the intersection curve at $\mathbf{p}$, let $\mathbf{v}$ be a unit vector parallel to L, and let $\mathbf{Q}_+$ be the "positive" side of L in $\mathbf{Q}$ consisting of all points $\mathbf{x} \in \mathbf{Q}$ such that

$$(\mathbf{x} - \mathbf{p}) \cdot \mathbf{N}(\mathbf{p}) \quad > \quad 0 \ .$$

If the curve through $\mathbf{p}$ determined by $\Sigma \cap \mathbf{Q}$ has positive normal curvature at $\mathbf{p}$, this means that the center of the osculating circle at $\mathbf{p}$ is a point of M that lies on $\mathbf{Q}_+$, and in fact it follows that all points on the curve that are close to $\mathbf{p}$, except for $\mathbf{p}$ itself, also lie on $\mathbf{Q}_+$. Similarly, let $\mathbf{Q}_-$ denote the other side of L in $\mathbf{Q}$ consisting of all points for which $(\mathbf{x} - \mathbf{p}) \cdot \mathbf{N}(\mathbf{p})$ is negative, and suppose that the Gaussian curvature of the curve through $\mathbf{p}$ determined by $\Sigma \cap \mathbf{Q}$ is negative. Then it follows that the center of the osculating circle at $\mathbf{p}$ is a point of M that lies on $\mathbf{Q}_-$, and in fact it follows that all points on the curve that are close to $\mathbf{p}$, except for $\mathbf{p}$ itself, also lie on $\mathbf{Q}_-$.

Suppose now that the Gaussian curvature at $\mathbf{p}$ is positive. This means that the maximum and minimum values for the normal curvatures of the intersection curves are nonzero and have the same sign. In this case it follows that all points of the intersection curves that are sufficiently close to $\mathbf{p}$ lie on one of the closed sides of the tangent plane that are determined by one of the inequalities $(\mathbf{x} - \mathbf{p}) \cdot \mathbf{N}(\mathbf{p}) \geq 0$ or $(\mathbf{x} - \mathbf{p}) \cdot \mathbf{N}(\mathbf{p}) \leq 0$. Furthermore, with the exception of $\mathbf{p}$ itself, all of the nearby points on such curves lie on the open sides defined by replacing $\leq$ and $\geq$ with strict inequalities. This corresponds to the notion of strict local convexity that was discussed in the exercises.

Before proceeding, it will be useful to set some notation. Given a nonzero tangent vector $\mathbf{w}$ at $\mathbf{p}$, let $\mathbf{B}(\mathbf{w})$ be equal to $\mathbf{w} \times \mathbf{N}(\mathbf{p})$; since $\mathbf{w}$ and $\mathbf{N}(\mathbf{p})$ are nonzero vectors that are perpendicular to each other, it follows that their cross product is nonzero and perpendicular to both of these vectors. The plane containing $\mathbf{p}$ with normal direction corresponding to $\mathbf{B}(\mathbf{w})$ will be denoted by $\mathbf{Q}(\mathbf{w})$, and its two open sides $\mathbf{Q}_\pm(\mathbf{w})$ may then be defined as in the previous paragraph.

Suppose now that the Gaussian curvature at $\mathbf{p}$ is negative. In this case one can choose nonzero tangent vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ at $\mathbf{p}$ such that the normal curvatures of the corresponding plane intersections $\mathbf{Q}(\mathbf{v}_1) \cap \Sigma$ and $\mathbf{Q}(\mathbf{v}_2) \cap \Sigma$ are positive in the first case and negative in the second. If β_i is the curve near $\mathbf{p}$ determined by the intersection $\mathbf{Q}(\mathbf{v}_i) \cap \Sigma$, then the centers of the osculating circles for β_1 and β_2 lie on $\mathbf{Q}(\mathbf{v}_1)_+$ and $\mathbf{Q}(\mathbf{v}_2)_-$ respectively, and similar statements hold for all points of β_1 and β_2 that are close to $\mathbf{p}$ except for $\mathbf{p}$ itself. A good model for this is the saddle surface defined by the equation

$$z \quad = \quad y^2 \ - \ x^2$$

near the origin. The tangent plane of this surface at the origin is the xy-plane, and the standard upward normal for the surface at the origin is the vector $(0, 0, 1)$. Direct calculation shows that the First Fundamental Form is

$$(1 \ + \ 4\,x^2)\, dx\, dx \ + \ 16\,x^2\,y^2 \ + \ (1 \ + \ 4\,y^2)\, dy\, dy \ .$$

Furthermore, the intersections of this surface with the yz- and xz-planes are the parabolas $z = y^2$ and $z = -x^2$ respectively. Away from the origin, these intersection curves lie in the open half planes determined by the strict inequalities $z > 0$ and $z < 0$ respectively. Incidentally, one can show directly that the saddle surface has negative Gaussian curvature at the origin using the formulas from Section IV.2 and this section as follows: By the formulas at the end of Section IV.2 we know that the Second Fundamental form is given by

$$\frac{2\,(dy\, dy \ - \ dx\, dx)}{\sqrt{1 \ + \ 4\,x^2 \ + \ 4\,y^2}}$$

and therefore the formula for the Gaussian curvature in terms of the Second Fundamental Form shows that

$$K \;\; = \;\; \frac{-4}{(1 \; + \; 4\,x^2 \; + \; 4\,y^2)^2} \; .$$

Thus the Gaussian curvature is negative at all points of this surface.

Note that the mean curvature must be nonzero if the Gaussian curvature is positive because the latter implies that the maximum and minimum values of the sectional curvature are both positive or both negative.

If the Gaussian curvature at $\mathbf{p}$ is zero, then one cannot draw many conclusions about the shape of the surface near $\mathbf{p}$. This is best seen using examples. There are two basic cases depending on whether the Second Fundamental Form is zero or nonzero. We shall begin by considering the second possibility. In this case the map $D\,\mathbf{N}(\mathbf{p})$ is not invertible but also nonzero, and therefore it has two eigenvalues, one of which is zero and one of which is nonzero. The cylinder defined by $x^2+y^2 = 1$ is one example of this sort. A typical point of this surface is the unit vector $(1,0,0)$, and its tangent plane is defined by the equation $x = 1$. In this case all the curves formed by intersecting the surface with planes containing the normal line at $(1,0,0)$, which is the x-axis, lie on the sides of the tangent plane determined by the inequality $x \leq 1$, and in fact all points on these curves except $(1,0,0)$ itself lie in the set determined by the strict inequality $x > 1$. However, it is also possible to describe other examples where $D\,\mathbf{N}(\mathbf{p})$ has rank 1 but the surface has points on both sides of the tangent plane. The graph of the function $z = x^2 + y^3$ at the origin is a specific example (look at the intersection with the yz-plane).

If the Gaussian curvature is zero and $D\,\mathbf{N}(\mathbf{p}) = 0$ then the mean curvature is also zero and the local behavior of the surface near $\mathbf{p}$ also cannot be determined without additional information. A plane is the simplest example of this type. However, there are also examples for which the surface is strictly locally convex near the point $\mathbf{p}$ and examples where the surface has points on both open sides of the tangent plane near $\mathbf{p}$. An example where strict local convexity holds is given by the graph of $f(x,y) = x^4 + y^4$ at the origin, where the tangent plane to the surface is merely the xy-plane. One can use the methods employed for the saddle surface to show that the Second Fundamental Form is zero at the origin. If one intersects this surface with a plane containing the normal line at the origin, which is the z-axis, then the resulting curves all lie on the side of the tangent plane defined by the inequality $z \geq 0$, and except for the origin itself all points of the curve lie on the open side where the strict inequality $z > 0$ holds. On the other hand, consider the *Monkey Saddle Surface* defined by the equation $z \;\; = \;\; x^3 \; - \; 3\,x^2\,y$ at the origin. Once again the tangent plane is the xy-plane and the Second Fundamental Form is zero. Using cylindrical coordinates and simple trigonometric identities, one can rewrite the equation of the surface as $z = r^3 \cos 3\theta$, and from this one sees that the intersection of the surface with a plane containing a normal line has a parametrization of the form $\left(t\,\cos 3\,\theta_0,\, t\,\sin 3\,\theta_0,\, t^3\right)$ for some fixed real number θ_0. These curves are line in the xy-plane if θ_0 is an integral multiple of $\pi/3$. On the other hand, for other choices of θ_0 the points on this curve corresponding to $t > 0$ and $t < 0$ lie on the two opposite open sides of the tangent plane defined by $z > 0$ and $z < 0$. In some cases the points corresponding to parameter values $t > 0$ lie on the side defined by $z > 0$, while in other cases these points lie on the side defined by $z < 0$. The following online sites contain excellent (and in the second case interactive) pictures of the Monkey Saddle:

`http://astronomy.swin.edu.au/~pbourke/surfaces/monkey/`

`http://www.ma.umist.ac.uk/kd/geomview/monkeysad.html`

`http://www.ag.jku.at/digpics_en.html`

IV.5 : Special classes of surfaces

(Lipschutz, Chapters 8–9)

In these notes particular attention has been given to understanding the main concepts in the differential geometry of surfaces for the objects encountered in analytic geometry and calculus, including quadric surfaces, surfaces of revolution and certain examples of ruled surfaces. Needless to say, mathematicians and scientists in related fields have also found numerous other examples of surfaces that are curious, interesting or important for one reason or another. The purpose of this section is to discuss a few additional examples beyond the usual ones from analytic geometry and calculus and also to comment further on the geometric interpretation of mean and Gaussian curvature for some standard examples that have not yet been considered. These and other examples are particularly useful in illustrating the sorts of geometric insights one can obtain by means of methods from ordinary and multivariable calculus.

Here are some online references that have particularly good collections of surface graphics:

`http://www.uib.no/People/nfytn/mathgal.htm`

`http://www.uta.edu/optics/sudduth/4d/the_main_gallery.htm`

`http://mathworld.wolfram.com/SurfaceofRevolution.html`

`http://www.math.arizona.edu/~models/Ruled_Surfaces/`

Of course, there are also many very good illustrations in O'Neill and do Carmo, but the impact advances in computer technology since the publication of the first editions of these books is clear (and the previously cited book by Gray goes into the uses of such technology for differential geometry in great detail).

Ruled surfaces

We begin with a very simple observation.

GAUSSIAN CURVATURE OF A RULED SURFACE. *If the surface Σ is given by a ruled parametrization in the sense of Section* III.2*, then its Gaussian curvature is nonpositive.*

The cylinder and plane are examples of ruled surfaces for which the Gaussian curvature is identically zero, and both the hyperboloid of one sheet and the hyperbolic paraboloid (saddle surface) are examples of ruled surfaces for which the Gaussian curvature is always negative.

Proof. We can do this without computing the Gaussian curvature explicitly. Suppose that we have a ruled parametrization

$$\mathbf{X}(u,v) \quad = \quad \mathbf{a}(u) + v \cdot \mathbf{b}(u)$$

where $\mathbf{a}'(u)$ is never zero and the vectors $\mathbf{a}'(u)$ and $\mathbf{b}(u)$ are always linearly independent. Then the space of tangent vectors at $\mathbf{X}(u,v)$ is spanned by the linearly independent vectors $\mathbf{a}'(u)$ and $\mathbf{b}(u)$.

Consider the curve through $\mathbf{p} = \mathbf{X}(u,v)$ formed by intersecting the tangent plane to $\mathbf{p}$ at that point with the unique plane that contains $\mathbf{X}(u,v)$ and whose normal line is parallel to the vector $\mathbf{N}(\mathbf{p}) \times \mathbf{b}(u)$. This intersection is locally given by the line through $\mathbf{p}$ that is parallel to $\mathbf{b}(u)$. Of course the curvature of this curve is equal to zero and therefore we know that there is one tangent direction at $\mathbf{p}$ for which the sectional curvature is zero. If the Gaussian curvature were either positive or negative, then the sectional curvature would be nonzero in every direction, and therefore the Gaussian curvature cannot be positive at $\mathbf{p}$.$\blacksquare$

Derivation of formulas for the mean and Gaussian curvature of a ruled surface are left to the reader as an exercise.

Ruled surfaces for which the Gaussian curvature is identically zero are called *developable surfaces*, and they have many significant properties. Further information on this topic may be found on pages 194 and 210 of do Carmo (also see pages 145–148 of D. V. Widder, *Advanced Calculus*, Second Edition, Dover, New York, 1989, ISBN: 0-486-66103-2).

Surfaces of revolution

We shall derive formulas for the Gaussian curvature of a surface of revolution obtained by rotating a curve in the xy-plane about the x- and y-axes. In the first case we need to assume that the x-coordinates for all points on the curve are positive, and in the second we need to make a similar assumption regarding the y-coordinates. Our ultimate goal is to describe a surface of revolution whose Gaussian curvature is equal to -1 at each point.

Before setting up the computations it is worthwhile to consider some examples in order to have a rough idea about what the general formulas for Gaussian curvature can be expected to yield. If we take $h(x) = \sqrt{1-x^2}$ and rotate it around the x-axis we obtain a portion of the unit sphere centered at the origin. This surface has Gaussian curvature equal to $+1$ at each point, and the second derivative of h is negative for $-1 < x < 1$. On the other hand, if we take $h(x) = \sqrt{1+x^2}$ and rotate it around the x-axis we obtain a portion of the hyperboloid of one sheet defined by the equation $y^2+z^2-x^2 = 1$, which has negative Gaussian curvature; in this case the second derivative of h is positive everywhere. This and further experimentation suggest that the signs of the second derivative and the Gaussian curvature should be the opposites of each other.

Suppose now that we are given a regular smooth curve $\mathbf{c}(t) = (\,p(t),\, q(t)\,)$ where $q(t) > 0$ for all t. Then the regular surface formed by rotating this curve about the x-axis may be given using the parametrization

$$\mathbf{X}(u,v) \quad = \quad \big(\,p(u),\; q(u)\,\cos v,\; p(u)\,\sin v\,\big)\ .$$

In order to compute the coefficients of the fundamental forms we need to find the partial derivatives $\mathbf{X}_1$ and $\mathbf{X}_2$ and the unit normal associated to the parametrization, which has the form

$$\mathbf{N} \quad = \quad \frac{1}{|\Omega|}\cdot\Omega$$

where $\Omega = \mathbf{X}_1 \times \mathbf{X}_2$. Here are the relevant formulas:

$$\begin{aligned}
\mathbf{X}_1 \quad &= \quad (\,p',\; q'\,\cos v,\; q'\,\sin v\,) \\
\mathbf{X}_2 \quad &= \quad (\,p,\; -q\,\sin v,\; q\,\cos v\,) \\
\Omega \quad &= \quad (\,q\,q',\; -p'\,q\,\cos v,\; -p'\,q\,\sin v\,) \\
|\Omega| \quad &= \quad q\,\sqrt{(p')^2 \;+\; (q')^2}
\end{aligned}$$

Given these formulas we see that the First Fundamental Form has coefficients $E = (p')^2 + (q')^2$, $F = 0$ and $G = q^2$.

Similarly, the second partial derivatives of $\mathbf{X}$ are given as follows:

$$\begin{aligned}
\mathbf{X}_{1,1} &= (\, p'', \, q'' \cos v, \, q'' \sin v \,) \\
\mathbf{X}_{1,2} &= \mathbf{X}_{2,1} \quad = \quad (\, 0, \, -q' \sin v, \, -q' \cos v \,) \\
\mathbf{X}_{2,2} &= (\, 0, \, -q \cos v, \, -q \sin v \,)
\end{aligned}$$

The corresponding inner products with Ω are given by

$$\begin{aligned}
\Omega \cdot \mathbf{X}_{1,1} &= p'' \, q \, q' \; - \; p' \, q \, q'' \\
\Omega \cdot \mathbf{X}_{1,2} &= 0 \\
\Omega \cdot \mathbf{X}_{2,2} &= p' \, q^2
\end{aligned}$$

and therefore the coefficients of the Second Fundamental Form are given as follows:

$$e \;=\; \frac{p'' \, q' \; - \; p' \, q''}{\sqrt{(p')^2 \; + \; (q')^2}}$$

$$f \;=\; 0$$

$$g \;=\; \frac{p' \, q}{\sqrt{(p')^2 \; + \; (q')^2}}$$

These computations yield the following formula for the Gaussian curvature:

$$K \;=\; \frac{e\,g - f^2}{E\,G - F^2} \;=\; \frac{e\,g}{E\,G} \;=$$

$$\frac{(\, p'' \, q \; - \; p' \, q'' \,) \, p' \, q}{[(p')^2 \; + \; (q')^2]^2 \, q^2} \;=\; \frac{(\, p'' \, q \; - \; p' \, q'' \,) \, p'}{[(p')^2 \; + \; (q')^2]^2 \, q}$$

SPECIAL CASES. Suppose first that $p(t) = t$ so that the curve is simply the graph of a smooth function. Then the formula reduces to

$$K \;=\; \frac{-q''}{[1 \; + \; (q')^2]^2 \cdot q}$$

and therefore the signs of K and q'' are opposite, exactly as our examples suggested.

Suppose now that we assume that $|\mathbf{c}'(t)| \equiv 1$, so that $(p')^2 \; + \; (q')^2 = 1$. If we differentiate this with respect to u we obtain the equation $p' \, p'' \; + \; q' \, q'' \;\; = \;\; 0$, and thus we may use these equations to rewrite the Gaussian curvature as $-q''/q$. When $\mathbf{c}$ gives the standard parametrization of the unit circle with $p(t) = \cos t$ and $q(t) = \sin t$, this gives another proof that the Gaussian curvature of the unit sphere is equal to 1, at least at all points except perhaps $(\pm 1, 0, 0)$.

Gaussian curvature of the torus. The preceding also allows us to compute the Gaussian curvature of the torus given given by revolving the circle with equation

$$x^2 \; + \; (y - 2)^2 \;\; = \;\; 1$$

about the x-axis. In this case parametric equations for the circle are given by $p(t) = \cos t$ and $q(t) = 2 + \sin t$, and by the formula given above the Gaussian curvature is equal to

$$\frac{-q''(t)}{q(t)} \;=\; \frac{\sin t}{2 \; + \; \sin t} \; .$$

This quantity is positive if $t \in (0, \pi)$, zero if $t = 0,\, \pi,\, 2\pi$, and negative if $t \in (\pi,\, 2\,\pi)$. Visually, it is positive on the piece of the surface obtained by revolving the upper semicircle about the x-axis, negative on the piece obtained by revolving the lower semicircle about the x-axis, and zero on the circles obtained by revolving the points $(\pm, 1,\, 1)$ about the x-axis. Note that the Second Fundamental Form is nonzero at all points where the Gaussian Curvature is equal to zero.

The tractrix and pseudosphere

We shall now apply the preceding calculations to find a surface of revolution whose Gaussian curvature is equal to a negative constant. The standard example of this sort is the **pseudosphere**, for which the curve one revolves around the x-axis is known as the **tractrix** that we shall now describe.

From a physical perspective the tractrix is given as follows: Suppose that a person is initially standing at the origin in $\mathbb{R}^2$ and is holding a tightly stretched leash with a dog on the other end at $(0, a)$ where $a > 0$. Now suppose that the person begins walking in the positive direction along the x-axis and the dog's path is such that the leash is tightly stretched at each point. If $\mathbf{D}(t)$ and $\mathbf{P}(t)$ denote the positions of the dog and person at time t, these conditions translate into the following mathematical conditions:

(1) The line of the leash is the tangent line of the path taken by the dog.

(2) The distance between the dog and person is always equal to a.

(3) If parametric equations for the dog's path are given by $(\, u(t),\, v(t)\,)$, then both $u(t)$ and $v(t)$ are positive while their derivatives satisfy $u'(t) \;\; > \;\; 0 \;\; > \;\; v'(t)$.

Here are some online graphics, including one that is animated:

`http://mathworld.wolfram.com/Tractrix.html`

`http://bradley.bradley.edu/∼delgado/122/Tractrix.pdf`

`http://www.amherst.edu/∼amcastro/MathMedia/galleries/Curves/Tractrix.html`

With the information given above we may derive parametric equations for the tractrix as follows: The position of the person $\mathbf{P}(t)$ on the x-axis is the intersection of that line with the tangent line, which may be parametrized as

$$\mathbf{L}(s) = \;=\; \mathbf{D}(t) + s\,\mathbf{D}'(t)$$

with the x-axis. If s is the parameter value at which this line meets the x-axis, then the mathematical conditions imply that $v(t) + s\,v'(t) = 0$, $|s\,\mathbf{c}'(t)| = a$ and $s\,u'(t) = \sqrt{a^2 - v(t)^2}$. Combining these equations, we conclude that

$$v(t)\,u'(t) \quad = \quad -s\,v'(t)\,u'(t) \quad = \quad -v'(t)\sqrt{a^2 - v(t)^2}\ .$$

Dividing these by the nonzero number $v(t)$ yields the differential equation

$$u' \quad = \quad \frac{\sqrt{a^2 - v^2} \cdot v'}{v}\ .$$

As usual, some initial conditions are needed in order to solve such equations uniquely. In our situation we know that

$$\lim_{t\to+\infty} u(t) \quad = \quad +\infty \qquad \text{and} \qquad \lim_{t\to+\infty} v(t) \quad = \quad +0 \; .$$

If we make the change of variables $y = a\,\sin\theta$ then standard antidifferentiation formulas from integral calculus show that

$$x \quad = \quad a\left(\ln\tan(\theta/2) \; + \; \cos\theta\right) \; + \; C$$

and since the limit of y as $\theta \to \frac{1}{2}\pi$ is equal to a, it follows that the limit of x as $\theta \to \frac{1}{2}\pi$ is equal to 0. If we substitue this into thre right hand side of the formula for x, we see that the constant of integration C must be equal to zero. This means that one can describe the tractrix analytically by means of the parametrization

$$\left(a\left(\ln\tan(\theta/2) \; + \; \cos\theta\right),\, a\,\sin\theta\right) .$$

Note that as θ goes from 0 to $\pi/2$ this traces out the curve in the **reverse** direction from the one considered originally; the limiting values at $\theta = 0$ and $\pi/2$ may be viewed as $(+\infty,\, 0)$ and $(0,\, a)$ respectively.

To find the Gaussian curvature of the pseudosphere we may now substitute the coordinates for this parametrization into the general formulas given before:

$$p(\theta) \quad = \quad a\left(\ln\tan(\theta/2) \; + \; \cos\theta\right)$$

$$q(\theta) \quad = \quad a\,\sin\theta$$

If one then simplifies the resulting expression using standard differentiation rules and trigonometric identities, the conclusion is that $K = -1/a^2$ at all points of the pseudosphere. Thus the latter is indeed the desired *surface of revolution with prescribed constant negative Gaussian curvature.*

Constant Gaussian curvature

Classical Euclidean and Noneuclidean geometry have a natural interpretation in differential geometry as spaces of constant curvature. We have already seen that the plane and the sphere are spaces that have constant curvature, with the constant value equal to zero in the planar case and a positive number in the spherical case. In classical geometry it is either implicitly or explicitly assumed that the spaces have translational symmetry — given two points one can find a rigid motion sending one point to the other. From the perspective of differential geometry, this corresponds to an assumption that the Gaussian curvatures at every pair of points are the same, or equivalently that the Gaussian curvature is constant. The structure of surfaces with constant Gaussian curvature has been a central topic in differential geometry throughout its history, and there are important results which imply that all surfaces with constant curvature are very closely related to the fundamental examples; namely, the plane in the case of zero curvature, the sphere in the case of positive curvature, and the Noneuclidean or Hyperbolic plane discussed at the end of Section III.4. A fundamental theorem of D. Hilbert shows that one does not have a nice realization of the latter as a surface in $\mathbb{R}^3$ (the details, which are beyond the scope of this course, are given in Section 5–11 of DO CARMO), but the pseudosphere provides a good model in $\mathbb{R}^3$ for a small portion of this object (more precisely, if one removes the copy of the tractrix corresponding to $v = \pi$, then the remaining portion of the pseudosphere is metrically equivalent to a region in the Hyperbolic Noneuclidean plane).

A surface is said to be a *minimal surface* if its mean curvature is identically zero. This condition is simple and analogous to the conditions for constant Gaussian curvature, but none of this explains the reason for the name. From the viewpoint of local differential geometry, a minimal surface is is one that is equally bent in all directions so as to have zero average curvature just like a plane or the surface $z = x^2 - y^2$ at the origin, but in contrast to the latter one wants this property at every point of the surface. Aside from the plane, two basic examples of such surfaces are the catenoid and helicoid that are discussed respectively on pages 244–245 and 219–220 of O'NEILL (see also pages 202–205 of DO CARMO).

Although the sign of the mean curvature of a surface depends upon the choice of unit normals, one can extend the concept of minimal surface to nonorientable surfaces because orientations exist locally and the vanishing property for the mean curvature is the same whether one starts with a given system of unit normals or their negatives.

Minimal surfaces are so named because of their connection to the following natural question: *Given a closed curve* Γ, *find the surface of least area that is bounded by* Γ. This question is known as *Plateau's Problem* and it is named after the physicist who noted that such surfaces may be realized physically by soap films that are bounded by the given curve.

As noted on page 197–199 of DO CARMO, surfaces of least area must have mean curvatures that are identically zero, and the discussion on those pages provides strong evidence for this, at least in some relatively elementary situations. A more detailed treatment appears in the following online document:

`http://ocw.mit.edu/NR/rdonlyres/Mathematics/` (*continue with next line*)
`18-994Fall-2004/179B4DC0-3C84-425A-93E1-9E9D06C83B0D/0/chapter11.pdf`

Despite the intuitive nature of the least area problem, a precise mathematical formulation of it in a reasonably general context turns out to be extremely nontrivial and requires methods beyond the scope of this course. However, there is a fairly accessible description of the key ideas in Chapter 18 of THORPE (see pages 156–160).

Minimal surfaces have important relations to the theory of functions of a complex variable and partial differential equations; the most basic aspects of this are described on pages 201–202 of DO CARMO. The study of minimal surfaces has had a strong impact on both geometry and analysis, in may cases leading to results on questions that at first doe not seem to have any relation to the least area problem.

Numerous examples of minimal surfaces have been discovered or constructed over the past two hundred years, and advances in computer technology during the past quarter century have led to striking new insights, yielding unexpected new types of such surfaces whose existence was first suggested by computer graphics and later confirmed by rigorous mathematical proofs (but not all potential examples arising from computer graphics turned out to be minimal surfaces!). A substantial amount of this work was motivated by potential applications of minimal surfaces to the other sciences and engineering. Here are some online references that discuss minimal surfaces, with many illustrations and more information on advances that have taken place during the past three decades:

`http://ctouron.freeshell.net/personal/costa/background.html`

`http://mathworld.wolfram.com/CostaMinimalSurface.html` http://mathworld.wolfram.com/MinimalSurface.html

http://www.indiana.edu/ minimal/toc.html

http://www.zib.de/polthier/booklet/intro.html

http://www.csuohio.edu/math/oprea/soap/soap.html

http://www.math.unifi.it/ paolini/diletto/minime/index.en.html

http://www.miqel.com/pure-math-patterns/visual-math-minimal-surfaces.html

The text by J. Oprea, *Differential Geometry and Its Applications*, Second Edition (Prentice-Hall, 2003, ISBN: 0-13-065246-6), contains a detailed and current account of minimal surfaces at the undergraduate textbook level.

V. Further Topics

This continuation of the course lecture notes discusses two fundamental topics in the classical theory of surfaces. The first is a basic result of Gauss which states that the Gaussian curvature of a surface depends only on the First Fundamental Form of the surface. This fact allows one to define curvature for a much broader range of geometric objects, an idea that has been basic to differential geometry for nearly two centuries. The second topic is an analog of the basic existence and uniqueness results for curves based upon the Frenet-Serret formulas. Just as curvature and torsion determine curves in 3-space, the First and Second Fundamental Forms provide a similarly complete characterization of surfaces.

V.1 : Compatibility equations, *Theorema Egregium*

(Lipschutz, Chapters 10–11)

One of the most far-reaching results on the differential geometry of surfaces is that the Gaussian curvature of a surface can be expressed entirely in terms of the First Fundamental Form:

GAUSS' THEOREMA EGREGIUM. *If* $\mathbf{X}$ *is a* $1-1$ *regular parametrization such that the First Fundamental Form is given by*

$$E(u,v)\,du\,du \;+\; 2\,F(u,v)\,du\,dv \;+\; G(u,v)\,dv\,dv$$

and $K(u,v)$ *is the Gaussian curvature function, then the Gaussian curvature depends only upon the coefficients of the First Fundamental Form of the surface and their partial derivatives.*

In contrast, the plane and cylinder have the same First Fundamental Form but different mean curvatures.

At the end of Section III.4 in the course lecture notes, we discussed generalizations of the First Fundamental Form known as Riemannian metrics. One can use the formula above to define Gaussian curvature with respect to an arbitrary Riemannian metric regardless of whether it comes from a First Fundamental Form. This is an important step in formulating general notion of curvature in differential geometry that can be used in many different contexts and have a dramatically wide range of applications in mathematics and physics.

Intrinsic geometry of surfaces

One way of interpreting Gauss' *Theorema Egregium* is to say that the Gaussian curvature is an **intrinsic** property of a surface; to quote from the Preface to O'Neill, such properties concern

> the geometry of a surface, *as seen by its inhabitants*, with no assumptions that the surface can be found in ordinary 3-dimensional space [or the manners in which it might be realized in ordinary 3-dimensional space].

Discussion of the proof

The results of this section and the next are based upon an analysis of the second partial derivatives of a regular parametrization $\mathbf{X}$. In some sense this is analogous to the idea behind the Frenet-Serret Formulas for curves; one writes out the various derivatives as linear combinations of simpler objects and looks for useful interrelationships. In the case of curves, the Frenet Trihedron provided a useful basis for $\mathbb{R}^3$ at each point of the curve. For surfaces given by regular parametrizations, the corresponding useful basis is given by the partial derivatives $\mathbf{X}_1$ and $\mathbf{X}_2$ together with the unit normal vector $\mathbf{N}$, which may be viewed as $\mathbf{X}_1 \times \mathbf{X}_2$ normalized to have unit length. One major difference with the theory for curves is that these bases are usually not orthonormal, but this turns out to be a relatively minor issue that can be addressed directly using linear algebra as in the final portion of Section IV.3 of the course lecture notes.

If one writes out the partial derivatives of $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{N}$ with respect to the u (first) and v (second) variables and uses the earlier computations involving the First and Second Fundamental Forms, one obtains the following sorts of formulas in which the quantities $\Gamma^i_{j,k}$ are smooth functions of u and v and are called *Christoffel symbols of the second kind*; the terminology is chosen to be consistent with concepts in tensor analysis (see the bottom of page 213 in the Schaum's Outline Series book on differential geometry for further information).

$$
\begin{aligned}
\mathbf{X}_{1,1} &= \Gamma^1_{1,1}\,\mathbf{X}_1 + \Gamma^2_{1,1}\,\mathbf{X}_2 + e\,\mathbf{N} \\
\mathbf{X}_{1,2} &= \Gamma^1_{1,2}\,\mathbf{X}_1 + \Gamma^2_{1,2}\,\mathbf{X}_2 + f\,\mathbf{N} \\
\mathbf{X}_{2,2} &= \Gamma^1_{2,2}\,\mathbf{X}_1 + \Gamma^2_{2,2}\,\mathbf{X}_2 + g\,\mathbf{N} \\
\mathbf{N}_1 &= \beta^1_1\,\mathbf{X}_1 + \beta^2_1\,\mathbf{X}_2 \\
\mathbf{N}_2 &= \beta^1_2\,\mathbf{X}_1 + \beta^2_2\,\mathbf{X}_2
\end{aligned}
$$

It is convenient to define $\Gamma^i_{2,1} = \Gamma^i_{1,2}$ for $i = 1,\, 2$ so that $\Gamma^i_{j,k}$ is defined for $1 \leq i,\, j,\, k \leq 2$ and satisfies $\Gamma^i_{k,j} = \Gamma^i_{j,k}$.

Using the methods described in the last part of Section IV.3 in the course lecture notes, one can solve for β^i_j in terms of the coefficients of the First and Second Fundamental Forms:

$$
\begin{aligned}
\beta_{1,1} &= \frac{f\,F - e\,G}{E\,G - F^2} \\
\beta_{2,1} &= \frac{e\,F - f\,E}{E\,G - F^2} \\
\beta_{1,2} &= \frac{f\,F - f\,G}{E\,G - F^2} \\
\beta_{1,2} &= \frac{f\,F - g\,E}{E\,G - F^2}
\end{aligned}
$$

If one substitutes these into the equations for $\mathbf{N}_1$ and $\mathbf{N}_2$ one obtains the *Weingarten equations*. Computing the Christoffel symbols is more difficult. The following formulas are derived in Problem 10.3 on page 216 of the Schaum's Outline Series review of differential geometry that was cited previously:

$$
\Gamma^1_{1,1} = \frac{G\,E_1 - 2\,F\,F_1 + F\,E_2}{2\,(E\,G - F^2)}
$$

$$\Gamma^1_{1,2} \quad = \quad \frac{G\,E_2 \; - \; F\,G_1}{2\,(E\,G \; - \; F^2)}$$

$$\Gamma^1_{2,2} \quad = \quad \frac{2\,G\,F_2 \; - \; G\,G_1 \; + \; F\,G_2}{2\,(E\,G \; - \; F^2)}$$

$$\Gamma^2_{1,1} \quad = \quad \frac{2\,E\,F_1 \; - \; E\,E_2 \; + \; F\,E_1}{2\,(E\,G \; - \; F^2)}$$

$$\Gamma^2_{1,2} \quad = \quad \frac{E\,G_1 \; - \; F\,E_2}{2\,(E\,G \; - \; F^2)}$$

$$\Gamma^2_{2,2} \quad = \quad \frac{E\,G_1 \; - \; 2\,F\,F_2 \; + \; F\,G_1}{2\,(E\,G \; - \; F^2)}$$

It is important to note that *the Christoffel symbols depend only upon the coefficients of the First Fundamental Form and their first partial derivatives.*

The most direct approach to proving Gauss' theorem about the Gaussian curvature is to continue by proving that

$$K\,(E\,G \; - \; F^2)^2 \quad = \quad [\mathbf{X}_{1,1},\, \mathbf{X}_1,\, \mathbf{X}_2\,] \cdot [\mathbf{X}_{2,2},\, \mathbf{X}_1,\, \mathbf{X}_2\,] \; - \; [\mathbf{X}_{1,2},\, \mathbf{X}_1,\, \mathbf{X}_2\,]^2 \; .$$

This compuation is carried out in Problem 10.4 on page 217 of the Schaum's Outline Series on differential geometry, and equivalent statements involving differential forms are established in Chapter 6 of O'NEILL (see pages 280–281 in particular). Further computations using the same methods then yield the identity

$$K\,(E\,G \; - \; F^2)^2 \quad = \quad \left(F_{1,2} \; - \; \tfrac{1}{2}\,E_{2,2} \; - \; \tfrac{1}{2}\,G_{1,1}\right) \cdot (E\,G \; - \; F^2) \; +$$

$$\begin{vmatrix} 0 & F_2 - \frac{1}{2}\,G_1 & \frac{1}{2}\,G_2 \\ \frac{1}{2}\,E_1 & E & F \\ F_1 - \frac{1}{2}\,E_2 & F & G \end{vmatrix} \; - \; \begin{vmatrix} 0 & \frac{1}{2}\,E_2 & \frac{1}{2}\,G_1 \\ \frac{1}{2}\,E_2 & E & F \\ \frac{1}{2}\,G_1 & F & G \end{vmatrix}$$

which implies that K depends only upon the coefficients of the First Fundamental Form and their partial derivatives. It is an elementary exercise in partial differentiation to show that this equation is equivalent to the one in the statement of Gauss' theorem that is given above.■

Another approach to deriving Gauss' theorem is given on pages 231–235 in Section 4–3 of DO CARMO and (using differential forms) in Chapter 6 of O'Neill. This alternate approach also has other implications, and it will be discussed in the next section.

Curvature and the First Fundamental Form

We have discussed the geometric significance of Gaussian curvature for a surface in $\mathbb{R}^3$ in terms of its First and Second Fundamental Forms. The **Theorema Egregium** provides a way of defining the Gaussian curvature entirely in terms of the First Fundamental Form, and consequently for riemannian metrics that are not necessarily realizable by surfaces in $\mathbb{R}^3$. One is therefore led to natural questions about interpreting the Gaussian curvature entirely in terms of metrical properties directly given the First Fundamental Form without using auxiliary objects such as normal lines or osculating circles. We shall describe one interpretation of positive and negative Gaussian curvature at a point entirely in metric terms; if the Gaussian curvature is equal to zero the situation is more complicated, but if the Gaussian curvature is identically zero then we shall give a similar interpretation.

Given a Riemannian metric

$$E(u,v)\,du\,du \;+\; 2\,F(u,v)\,du\,dv \;+\; G(u,v)\,dv\,dv$$

and a parametrized regular, piecewise smooth curve in a connected domain $U \subset \mathbb{R}^2$ on which the metric is defined, one can define the **length** of the curve by the formula

$$\int_a^b \sqrt{\,E(u,v)\,u'(t)^2 \;+\; 2\,F(u,v)\,u'(t)\,v'(t) \;+\; G(u,v)\,v'(t)^2}\;dt$$

where the curve is defined on the interval $[a,b]$. The positivity condition on the coefficients E, F and G for a Riemannian metric imply that the expression inside the square root sign is always positive for regular smooth curves. One would like to define the *distance between two points* with respect to this metric as the greatest lower bound of the lengths of all regular piecewise smooth curves joining the points.

Two questions immediately arise. First of all, one needs to show that the lengths of curves joining two distinct points are bounded from below by a positive constant; in other words, if $\mathbf{p}$ and $\mathbf{q}$ are distinct points of a surface then it is not possible to find a sequence of piecewise smooth regular curves $\mathbf{y}_n$ joining them such that the length of $\mathbf{y}_n$ is less than $1/n$. Second, one would like to know if there is some curve for which the greatest lower bound is actually realized. Such a curve is called a *minimal geodesic*.

It is fairly easy to construct a somewhat artificial example where there is no curve of minimum length joining two points. Specifically, consider the surface given by removing the origin from the xy-plane. Then the greatest lower bound of the lengths of all piecewise smooth curves joining $(1,0,0)$ and $(-1,0,0)$ is equal to 2, which is the ordinary Euclidean distance, but there is no curve of length 2 joining these points that misses the origin. To see this, let $\mathbf{y}$ be a regular piecewise smooth curve joining the two points in question that is defined on $[a,b]$. Then there is some point $\xi \in (a,b)$ such that the second coordinate of $\mathbf{y}(\xi)$ is equal to zero; it follows the first coordinate of $\mathbf{y}(\xi)$ must be equal to some nonzero value, say c. This in turn implies that the arc length of $\mathbf{y}$ must be greater than or equal to the length of the broken line curve which first joins $(1,0,0)$ to $(0,c,0)$ linearly and then joings $(0,c,0)$ to $(-1,0,0)$ linearly. The length of this broken line curve is $2\sqrt{1+c^2}$, which is strictly greater than 2. Therefore there is no curve of shortest length joining the two points that lies completely inside the surface. One obvious feature of this example is that one can extend the given surface to a larger one (namely, the whole plane) in which there is a curve of minimum length joining the two points in question. In fact, one can construct examples for which one cannot add extra points to ensure that minimizing geodesics always exist, but such a construction would require a great deal of additional work. A natural candidate for a bad example is the graph of the function

$$f(x,y) \quad = \quad \frac{x\,y}{x^2 \;+\; y^2}$$

which is defined for $(x,y) \neq (0,0)$ and cannot be extended to a function that is continuous at $(0,0)$.

In contrast to the preceding paragraph, it turns out that one can always find curves of minimum length joining a given point $\mathbf{p}$ to another point $\mathbf{q}$ provided $\mathbf{q}$ is sufficiently close to $\mathbf{p}$, and this fact has important implications to showing the the lengths of curves joining two distinct points are bounded from below by a positive constant.

EXISTENCE OF SHORT GEODESICS. *Suppose we are given a riemannian metric* $\mathbf{M}$ *on a connected domain in* $\mathbb{R}^2$, *and let* $\mathbf{p} \in U$. *Then there is an* $r > 0$ *such that* $|\mathbf{q} - \mathbf{p}| < r$ *implies that* $\mathbf{p}$ *and* $\mathbf{q}$ *can be joined by a regular piecewise smooth curve of least length, and this curve is in fact a regular smooth curve that lies entirely in the open disk with center* $\mathbf{p}$ *and radius* r.

Furthermore, given any nonzero vector $\mathbf{v} \in \mathbb{R}^2$ *there is a unique regular smooth curve curve* Γ *defined on an open interval* $(-h, h)$ *containing* 0 *such that* $\Gamma(0) = \mathbf{p}$, $\Gamma'(0) = \mathbf{v}$ *and* Γ *defines a curve of minimum length joining* $\mathbf{p}$ *to* $\Gamma(t)$ *for all* t *in the given interval* $(-h, h)$.

Finally, if $\delta \in (0, r)$ *and* $L(\mathbf{q})$ *denotes the length of the shortest curve joining* $\mathbf{p}$ *to* $\mathbf{q}$, *then the minimum value* $m(\delta)$ *of* $L(\mathbf{q})$ *over the circle defined by* $|\mathbf{q}| = \delta$ *is positive.*■

The curves of least length in this result are called **minimizing geodesics**. It turns out that such curves are defined by second order differential equations, and this is the reason for the conclusion in the second paragraph. At the end of this section we shall include a few remarks on geodesics for oriented surfaces in $\mathbb{R}^3$.

COROLLARY. *If* Σ *is a surface and* $\mathbf{p}$ *and* $\mathbf{q}$ *are two points on* σ *that can be joined by a regular piecewise smooth curve on* Σ, *then the set of lengths for all such curves is bounded from below by a positive constant.*

Proof. To simplify the discussion we shall choose parametrizations for our regular piecewise smooth curves over some interval of the form $[0, a]$ such that $\Gamma(0) = \mathbf{p}$ and $\Gamma(a) = \mathbf{q}$. We need to find a positive lower bound for the length that does not depend upon the particular curve Γ.

Let $\mathbf{X}$ be a regular smooth parametrization for Σ at $\mathbf{p}$ that is 1–1, let $\mathbf{X}(\mathbf{p}_0) = \mathbf{p}$, and let $r > 0$ be as in the existence theorem stated above. There are two cases, depending upon whether the point $\mathbf{q} \in \Sigma$ has the form $\mathbf{X}(\mathbf{q}_0)$ for some $\mathbf{q}_0$ satisfying $|\mathbf{q}_0 - \mathbf{p}_0| < r$.

First Case. Suppose that $\mathbf{q}$ satisfies the condition in the preceding sentence, and let $s = |\mathbf{q}_0 - \mathbf{p}_0|$. If $\mathbf{y}$ is an arbitrary point of Σ having the form $\mathbf{X}(\mathbf{y}_0)$ for some $\mathbf{y}_0$ satisfying $|\mathbf{y}_0 - \mathbf{p}_0| < r$, then we shall define $g_0(\mathbf{y})$ to be equal to $|\mathbf{y}_0 - \mathbf{p}_0|$; the right hand side is well defined because the parametrization $\mathbf{X}$ is 1–1. This turns out to be a continuous function of $\mathbf{y}$. Likewise, if we define a real valued function g by setting $g(t) = \min\{s,\, g_0(\,\Gamma(t)\,)\,\}$ if $\Gamma(t)$ has the given special form, and $g(t) = s$ if $\Gamma(t)$ does not have this form, then g is continuous on the interval $[a, b]$ over which Γ is defined.

Since $g(a) = 0$ and $g(b) = s$, there must be a first parameter value t_0 such that $g(t_0) = s$. We claim that the image of the restricted curve $\Gamma|[0, t_0)$ lies in the image W of the disk of radius s centered at $\mathbf{p}_0$ under $\mathbf{X}$ (in fact a stronger statement is true but we shall not need this). This is true because if $\Gamma(t)$ does not lie in the image then $g(t) \geq s$ and we know that $g(t) < s$ if $t \in [0, t_0)$.

By the Intermediate Value Theorem there is a $t_1 \in (0, t_0)$ such that $g(t_1) = \frac{1}{2}s$; we know that the image of Γ restricted to $[0, t_1]$ lies in the set W described above, and therefore this restriction may be written as a composite $\mathbf{X} \circ \Gamma_1$ for some regular piecewise smooth curve Γ_1 which takes values in the disk of radius r centered at $\mathbf{p}_0$. We then have

$$\text{Length}\,(\Gamma|[0, t_1]) \quad = \quad \text{Length}_{\mathbf{M}}(\Gamma_1) \geq m(\tfrac{1}{2}\,s) \quad > \quad 0$$

on one hand and

$$\text{Length}\,(\Gamma|[0, t_1]) \quad \leq \quad \text{Length}(\Gamma)$$

on the other, which implies that the right had side is greater than or equal to the positive quantity $m(\frac{1}{2}s)$. This gives us our desired positive lower bound on the length of Γ which is independent of the curve Γ itself.

SECOND CASE. The argument is similar but not quite identical. We may define the function g exactly in the first case for an arbitrary s such that $0 < s < r$. In this case we know that $g(t) = s$ for some parameter value t because there is some value t such that $\Gamma(t)$ does **NOT** have the form $\mathbf{X}(\mathbf{y}_0)$ for some $\mathbf{y}_0$ satisfying $|\mathbf{y}_0 - \mathbf{p}_0| < r$. We can now proceed as before to find the least parameter value t_0 such that $g(t_0) = s$, and from this point on the argument is identical to the proof in the first case.■

FUNDAMENTAL PROPERTIES OF DISTANCE FUNCTIONS. *Suppose that we have either a Riemannian metric $\mathbf{M}$ defined on a connected domain U in $\mathbb{R}^n$ or a geometric surface Σ in $\mathbb{R}^3$ such that each pair of points in Σ can be joined by a regular piecewise smooth curve in Σ, and let $d_{\mathbf{M}}(\mathbf{x},\, \mathbf{y})$ or $d_\Sigma(\mathbf{x},\, \mathbf{y})$ denote the greatest lower bound of the lengths of piecewise smooth curves joining $\mathbf{x}$ and $\mathbf{y}$ in U or Σ. Then this distance function d has the following basic properties:*

[1] *The distance $d(\mathbf{x},\, \mathbf{y})$ is nonnegative, and it is equal to zero if and only if $\mathbf{x} = \mathbf{y}$.*

[2] *For all $\mathbf{x}$ and $\mathbf{y}$ we have $d(\mathbf{x},\, \mathbf{y}) = d(\mathbf{y},\, \mathbf{x})$. and it is equal to zero if and only if $\mathbf{x} = \mathbf{y}$.*

[3] (TRIANGLE INEQUALITY) *For all $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ we have*

$$d(\mathbf{x},\, \mathbf{z}) \quad \leq \quad d(\mathbf{x},\, \mathbf{y}) \ + \ d(\mathbf{y},\, \mathbf{z})$$

Sketch of proofs. The first statement follows from the immediately preceding discussion. To prove the second, not that if Γ is a regular piecewise smooth curve defined on $[a, b]$ joining $\mathbf{x}$ to $\mathbf{y}$ then $\Gamma^*(t) = \Gamma(b - t)$ defines a similar curve on $[a - b, 0]$ joining $\mathbf{y}$ to $\mathbf{x}$. This implies that $d(\mathbf{y},\, \mathbf{x}) \leq d(\mathbf{x},\, \mathbf{y})$. Reversing the roles of $\mathbf{x}$ and $\mathbf{y}$ yields the reverse inequality $d(\mathbf{y},\, \mathbf{x}) \geq d(\mathbf{x},\, \mathbf{y})$, and therefore the two quantities must be equal. Finally, to prove the third statement, let $\varepsilon > 0$ and choose suitable curves Γ_1 and Γ_2 such that Γ_i is defined on $[0, a_i]$, with Γ_1 joining $\mathbf{x}$ to $\mathbf{y}$ and Γ_2 joining $\mathbf{y}$ to $\mathbf{z}$, and the lengths of these curves satisfying

$$\text{Length}(\Gamma_1) \quad \leq \quad d(\mathbf{x},\, \mathbf{y}) \ + \ \frac{\varepsilon}{2}$$

$$\text{Length}(\Gamma_2) \quad \leq \quad d(\mathbf{y},\, \mathbf{z}) \ + \ \frac{\varepsilon}{2}\ .$$

Consider the curve formed by concatenating Γ_1 and Γ_2; specifically, let Γ be the curve defined on the interval $[0, a_1 + a_2]$ such that $\Gamma(t) = \Gamma_1(t)$ for $t \in [0, a_1]$ and $\Gamma(t) = \Gamma(t - a_1)$ for $t \in [a_1, a_1 + a_2]$. These piece together to form a regular piecewise smooth curve because the two formulas yield the same point at parameter value a_1. The length of this curve then given by

$$\text{Length}(\Gamma_1) \ + \ \text{Length}(\Gamma_2)$$

and hence we have the inequality

$$d((\mathbf{x},\, \mathbf{z}) \quad = \quad \text{Length}(\Gamma) \quad = \quad \text{Length}(\Gamma_1) \ + \ \text{Length}(\Gamma_2) \quad \leq \quad d(\mathbf{x},\, \mathbf{y}) \ + \ d(\mathbf{y},\, \mathbf{z}) \ + \ \varepsilon$$

for every $\varepsilon > 0$. In particular this implies that the expression on the left hand side cannot be greater than $d(\mathbf{x},\, \mathbf{y}) \ + \ d(\mathbf{y},\, \mathbf{z})$, and this is precisely the assertion in [3].■

A METRIC INTERPRETATION OF CURVATURE. Suppose that we are given three points $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ in $\mathbb{R}^3$ that form the vertices of an isosceles triangle with vertex at $\mathbf{a}$; *i.e.*, we have $|\mathbf{b} - \mathbf{a}| = |\mathbf{c} - \mathbf{a}| = \ell > 0$. If θ is the angle between $\mathbf{b} - \mathbf{a}$ and $\mathbf{c} - \mathbf{a}$ then it is an elementary exercise in trigonometry to prove that $|\mathbf{c} - \mathbf{b}| = 2\,\sin\frac{1}{2}\theta$. Roughly speaking, Gaussian curvature measures the extent to which this fails for riemannian metrics. The proof of this fact requires a considerable amount of machinery from Riemannian geometry, so we shall simply state the results here.

Since we are only concerned with metric behavior near a point, it will suffice to look at Riemannian metrics defined on an open disk centered at some point $\mathbf{p}$ in a connected domain $U \subset \mathbb{R}^2$. Let $\mathbf{M}$ be a Riemannian metric, and let $r > 0$ be so small that every point in the open disk of radius r centered at $\mathbf{p}$ can be joined to the later by a smooth curve of minimum length lying entirely inside this disk. Given two linearly independent vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^2$, let $\theta_{\mathbf{M}}(\mathbf{p})$ be the angle between them computed with respect to the Riemannian metric:

$$\cos\left(\theta_{\mathbf{M}}(\mathbf{p})\right) \quad = \frac{\mathbf{M}_{\mathbf{p}}(\mathbf{v},\, \mathbf{w}\,)}{\left(\mathbf{M}_{\mathbf{p}}(\mathbf{v},\, \mathbf{v}\,)\right)^{1/2}\, \left(\mathbf{M}_{\mathbf{p}}(\mathbf{w},\, \mathbf{w}\,)\right)^{1/2}}$$

Consider now the smooth geodesics which pass through $\mathbf{p}$ and have tangent vectors $\mathbf{v}$ and $\mathbf{w}$ at $\mathbf{p}$. We can find points on these geodesics that are some positive distances away from $\mathbf{p}$; if δ_0 is the minimum of the two distances, then for every $\ell \in (0, \delta_0]$ we can find points $\mathbf{x}$ and $\mathbf{y}$ on the respective geodesics such that the distances from $\mathbf{x}$ and $\mathbf{y}$ to $\mathbf{p}$ are both equal to ℓ (suppose that we have geodesics with the given tangent vectors defined on intervals $[0, a]$ and $[0, b]$ respectively; then by the Intermediate Value Theorem one can find points s_0 and t_0 in these intervals so that the lengths of the restrictions of the geodesics up to parameter values s_0 and t_0 are equal to ℓ). We then have the following relationships between the Gaussian curvature at $\mathbf{p}$ and the distance between $\mathbf{x}$ and $\mathbf{y}$ with respect to $\mathbf{M}$.

DISTANCE COMPARISON. *Suppose we are given everything as in the preceding discussion, and let K be the Gaussian curvature at $\mathbf{p}$.*

(i) If $K > 0$ then there is a $\delta_1 > 0$ such that if $\ell < \delta_1$ we have

$$d_{\mathbf{M}}(\mathbf{x},\, \mathbf{y}) \quad < \quad 2\,\ell\, \sin\theta_{\mathbf{M}}(\mathbf{p})\ .$$

(ii) If $K < 0$ then there is a $\delta_1 > 0$ such that if $\ell < \delta_1$ we have

$$d_{\mathbf{M}}(\mathbf{x},\, \mathbf{y}) \quad > \quad 2\,\ell\, \sin\theta_{\mathbf{M}}(\mathbf{p})\ .$$

(iii) If the Gaussian curvature is identically zero, then there is a $\delta_1 > 0$ such that if $\ell < \delta_1$ we have

$$d_{\mathbf{M}}(\mathbf{x},\, \mathbf{y}) \quad = \quad 2\,\ell\, \sin\theta_{\mathbf{M}}(\mathbf{p})\ . \blacksquare$$

These may be viewed as generalizations of standard trigonometric formulas from spherical geometry, Noneuclidean geometry in the sense of Bólyai and Lobachevsky, and classical Euclidean geometry respectively. Note that if we only know the Gaussian curvature is zero at $\mathbf{p}$ but we know nothing else about its behavior near $\mathbf{p}$, then these comparison results yield no information.

Equations defining geodesics

If we are given a surface in $\mathbb{R}^3$, then there is a very simple characterization of minimizing geodesics.

PROPOSITION. *Let S be an oriented surface in $\mathbb{R}^3$ with normal vector field $\mathbf{N}$. Then the following hold:*

(i) If $\gamma(s)$ is a minimizing geodesic in S such that $|\gamma'(s)| = 1$ for all s, then $\gamma''(s)$ is a scalar multiple of $\mathbf{N}\big(\gamma(s)\big)$ for all parameter values s, and hence $\gamma''(s)$ is perpendicular to the 2-dimensional subspace of tangent vectors to S at $\gamma(s)$.

(ii) Conversely, if $\gamma(s)$ is a curve in S such that $|\gamma'(s)| = 1$ for all s and $\gamma''(s)$ is a scalar multiple of $\mathbf{N}\big(\gamma(s)\big)$ for all parameter values s, then for each value s_0 there is a $\delta > 0$ such that if $0 < h < \delta$ then the restriction of γ to $[s_0 - h, s_0 + h]$ is a minimizing geodesic.■

More generally, a smooth curve γ as a above in an oriented surface S (with normal vector field $\mathbf{N}$) is said to be a **geodesic** if $|\gamma'(s)| = 1$ for all s, then $\gamma''(s)$ is a scalar multiple of $\mathbf{N}\big(\gamma(s)\big)$ for all parameter values s. Further information about such curves and proofs of the assertions above can be found on pages 232–238 of Schaum's Outline Series on Differential Geometry.

It is not difficult to check that the given condition holds for the two most important examples of minimizing geodesics. In particular, if S is a plane, then the condition reduces to saying that γ'' is always perpendicular to the plane; on the other hand, we can also check that γ'' must lie in the unique 2-dimensional subspace V which is equal or parallel to S, and hence the condition in the proposition reduces to $\gamma'' = \mathbf{0}$, which is the differential equation of a linear curve in $\mathbb{R}^3$. Similarly, if S is a sphere and we take a great circle arc, then γ'' is a scalar multiple of γ, and since the tangent plane's normal line is given by the radial line, it follows that γ'' is perpendicular to the 2-dimensional vector subspace of tangent vectors at a given point.

The spherical examples also show that not every geodesic in the general sense is a minimizing geodesic. By the preceding discussion we know that every great circle arc is a geodesic (use the condition on γ''), but if we are given two points $\mathbf{p}$ and $\mathbf{q}$ on the sphere that are not opposite each other (in other words, the line joining them does not pass through the center of the sphere), then the great circle splits into a major and minor arc, and the major arc cannot be a minimizing geodesic because it is longer than the minor arc.

V.2 : Fundamental Theorem of Local Surface Theory

(Lipschutz, Appendix II)

The Frenet-Serret Formulas imply that curvature and torsion completely determine a curve locally provided on gives the initial position and unit tangent vector for the curve. There is a corresponding theorem for surfaces involving the coefficients E, F, G and e, f, g of the First and Second Fundamental Forms. However, these coefficient functions must satisfy some nontrivial restrictions. We have already noted that the matrix for the First Fundamental Form

$$\begin{pmatrix} E(u,v) & F(u,v) \\ F(u,v) & G(u,v) \end{pmatrix}$$

must have positive eigenvalues, or equivalently that E and G as well as the determinant $EG - F^2$ must be positive. However, there are also other conditions that arise naturally from our basic assumptions that a local parametrization $\mathbf{X}$ have "sufficiently many" continuous partial derivatives. In particular, if we want $\mathbf{X}$ to have continuous third partial derivatives then we have equations of the form $\mathbf{X}_{1,1,2} = \mathbf{X}_{1,2,1} = \mathbf{X}_{2,1,1}$ and then we have equations of the form $\mathbf{X}_{2,2,1} = \mathbf{X}_{2,1,2} = \mathbf{X}_{1,2,2}$. If we combine these equations with the expansions of the second partial derivatives $\mathbf{X}_{i,j}$ in terms of

Christoffel symbols and the Second Fundamental Form coefficients, we obtain the following three equations:

$$e_2 - f_1 \;=\; e\,\Gamma^1_{1,2} \;+\; f\,\left(\Gamma^2_{1,2} \;-\; \Gamma^1_{1,1}\right) \;-\; g\,\Gamma^2_{1,1}$$

$$f_2 - g_1 \;=\; e\,\Gamma^1_{2,2} \;+\; f\,\left(\Gamma^2_{2,2} \;-\; \Gamma^1_{1,2}\right) \;-\; g\,\Gamma^2_{1,2}$$

$$e\,g \;-\; f^2 \;=\; F\cdot\left[(\Gamma^2_{2,2})_1 \;-\; (\Gamma^2_{1,2})_2 \;+\; \Gamma^1_{2,2}\Gamma^2_{1,1} \;-\; \Gamma^1_{1,2}\Gamma^2_{1,1}\right] \;+$$

$$E\cdot\left[(\Gamma^1_{2,2})_1 \;-\; (\Gamma^1_{1,2})_2 \;+\; \Gamma^1_{2,2}\Gamma^1_{1,1} \;+\; \Gamma^2_{2,2}\Gamma^1_{1,2} \;-\; \Gamma^1_{1,2}\Gamma^1_{1,2} \;-\; \Gamma^2_{1,2}\Gamma^1_{2,2}\right]$$

The first two of these are known as the *Codazzi-Mainardi Equations*. We note in passing that the third equation provides another demonstration of Gauss' Theorema Egregium; in fact, one important advantage of this proof is that it reflects the standard approach to curvature in the study of differential geometry for objects whose dimensions are greater than two.

The verifications of these formulas from the classical viewpoint are carried out on pages 235–236 of DO CARMO and in Problem 10.28 on page 224 of the Schaum's Outline Series book on differential geometry. Derivation of the corresponding formulas involving differential forms are given on pages 257, 260 and 281 of O'NEILL (see Theorem 1.7, Corollary 2.3 and Theorem 5.4 respectively).■

The Gauss and Codazzi-Mainardi equations play an important role in establishing the main result of this section.

FUNDAMENTAL THEOREM OF LOCAL SURFACE THEORY. *Let U be a connected domain in $\mathbb{R}^2$, and let E, F, G and e, f, g be smooth functions with sufficiently many continuous partial derivatives on U such that E, F and G satisfy the positive definiteness conditions given above and e, f and g satisfy the three compatibility conditions displayed above. Then for each $\mathbf{p}_0 \in U$, $\mathbf{p} \in \mathbb{R}^3$ and plane Π containing $\mathbf{p}$, there is a regular surface parametrization $\mathbf{X}$ defined on some open disk N about $\mathbf{p}_0$ such that the First and Second Fundamental Forms of $\mathbf{X}$ have coefficients equal to E, F, G and e, f, g respectively. This parametrization is locally unique up to a rigid motion of $\mathbb{R}^3$.*

The uniqueness proof is essentially a relatively lengthy argument involving the uniqueness of solutions of certain ordinary differential equations (see pages 236 and 311–314 of DO CARMO or the argument following the statement of Theorem 10.4 on pages 203–204 of the Schaum's Outline Series book on differential geometry). On the other hand, the existence proof requires the solution of a system of partial differential equations.

In order to prove the existence of a regular smooth surface parametrization it is necessary to solve partial differential equations of the form $D\,y = A(\,x, y\,)$ where x and y are vectors and A is a smooth matrix valued function of x and y. In contrast to the situation for ordinary differential equations, the partial differential equation given above does not necessarily have a solution; specifically, the standard mixed partial derivative identities

$$\frac{\partial^2}{\partial x_i\,\partial x_j} \;=\; \frac{\partial^2}{\partial x_j\,\partial x_i}$$

imply that the entries of $A(x, y)$ and their partial derivatives must satisfy certain equations. However, the following result of F. G. Frobenius ensures that solutions always exist provided these conditions are satisfied:

FROBENIUS INTEGRABILITY THEOREM. *Let $n = k + d$, identify $\mathbb{R}^n$ with $\mathbb{R}^k \times \mathbb{R}^d$, let U be a connected domain in $\mathbb{R}^n$ and Let $\mathbf{A}$ be a smooth function defined on U and taking values in the space of $d \times k$ matrices, and let $(\mathbf{a}.\,\mathbf{b}\,) \in U$. Denote the entries of $\mathbf{A}$ by $A_{i,j}$.*

Assume in addition that these functions satisfy the compatibility conditions

$$\frac{\partial A_{i,j}}{\partial x_r} + \sum_{s=1}^{d} \frac{\partial A_{i,j}}{\partial x_s} A_{s,r} = \frac{\partial A_{i,r}}{\partial x_j} + \sum_{s=1}^{d} \frac{\partial A_{i,r}}{\partial x_s} A_{s,j} .$$

Then there exists a unique function Φ *defined on an open disk* V *containing* $\mathbf{a}$ *and taking values in* $\mathbb{R}^d$ *such that the following conditions hold:*

[1] $\Phi(\mathbf{a}) = \mathbf{b}$

[2] $(\mathbf{x}, \Phi(\mathbf{x})) \in U$ *for all* $\mathbf{x} \in V$.

[3] $D\,\Phi(\mathbf{x}) = \mathbf{A}(\mathbf{x}, \Phi(\mathbf{x}))$

Conversely, if such a function exists then the compatibility condition is satisfied.

Biographical information on Frobenius, and also many other mathematicians, may be found at the following online site:

`http://www-gap.cds.st-and.ac.uk/~history/BiogIndex.html`

The proof of the existence portion of the Fundamental Theorem of Local Surface Theory is discussed on pages 311-314 of do Carmo as well as in Appendix 2 on pages 264–265 of the Schaum's Outline Series book on differential geometry.■

Final remarks

1. A discussion of the Fundamental Theorem of Local Surface Theory from the viewpoint in Chapter 6 of O'Neill appears in Section 6.9 of the latter, with the main result appearing as Theorem 9.2 on pages 306–307. As noted at the end of this discussion, the result is completely analogous to the congruence theorem for curves discussed in Section II.4 of the course lecture notes.

2. A generalization of the Fundamental Theorem of Local Surface Theory to hypersurfaces of dimension $(n-1)$ in $\mathbb{R}^n$ is established in Section 9.2 of Hicks; the argument is a direct generalization of the proof for surfaces.■

V.3 : Riemannian metrics and hyperbolic geometry

(Lipschutz, Chapter 11)

In Section III.4 of the lecture notes we mentioned that the non-Euclidean plane discovered independently by L. Bólyai, N. I. Lobachevsky and C. F. Gauss in the 19^{th} century has a natural interpretation in terms of riemannian metrics. Specifically, one takes the underlying space U to be the open unit disk about the origin in $\mathbb{R}^2$, and the riemannian metric given to the so-called **Poincaré disk metric:**

$$\frac{dx\,dx \ + \ dy\,dy}{(1-x^2-y^2)^2}$$

In the notes we described the curves of shortest length (with respect to this metric) that join pairs of points in U. The purpose here is to explain the connection between this object and non-Euclidean geometry in terms of congruence.

The classical geometric notion of congruence has two basic properties:

(1) Given points $\mathbf{x}$ and $\mathbf{y}$ in U, there is a rigid motion (which must be an isometry) taking $\mathbf{x}$ to $\mathbf{y}$.

(2) Given a point $\mathbf{p}$ in U and two unit tangent vectors $\mathbf{x}$ and $\mathbf{v}$ at $\mathbf{p}$ there is an isometry T which sends $\mathbf{p}$ to itself and sends the curve $\gamma(t) = \mathbf{p} + t\mathbf{x}$ to the curve $T \circ \gamma$ such that $(T \circ \gamma)'(0) = \mathbf{y}$.

The notions of isometry and unit vector should be interpreted within the framework of riemannian metrics, and in this connection we use the following characterization of (smooth) riemannian isometries:

PROPOSITION. *Suppose that we are given a riemannian metric g over a connected domain in $\mathbb{R}^2$; by construction, if $u \in U$ then this yields an inner product g_u on $\mathbb{R}^2$ such that the Gram matrix coefficients*

$$g_{i,j}(u) \quad = \quad g_u(\mathbf{e}_i,\, \mathbf{e}_j)$$

are smooth functions of u (continuous first partial derivatives at least). Suppose that $f : U \to U$ is a smooth $1 - 1$ onto map with a smooth inverse such that for all $(u, \mathbf{x}, \mathbf{y}) \in \mathbb{R}^n$ we have the following identity:

$$g_u(\mathbf{x}, \mathbf{y}) \quad = \quad g_{f(u)}\big(\, Df(u)\mathbf{x},\; Df(u)\mathbf{y}\,\big)$$

Then for each smooth curve γ from the closed interval $[a, b]$ to U, the lengths of γ and $f \circ \gamma$ are equal. In particular, if $\mathbf{p}$ and $\mathbf{q}$ are in U and γ is a smooth curve of shortest length joining $\mathbf{p}$ and $\mathbf{q}$, then $f \circ \gamma$ is a smooth curve of shortest length joining $f(\mathbf{p})$ and $f(\mathbf{q})$.

The equality of length follows directly from the isometry identity, and the statement about curves of shortest length follows because f preserves lengths of curves.∎

Definition. A map f satisfying the condition in the proposition will be called a *riemannian isometry.* If we are working with riemannian isometries, then the second condition involving congruence can be reformulated as follows:

(2′) Given a point $\mathbf{p}$ in U and two unit tangent vectors $\mathbf{x}$ and $\mathbf{y}$ at $\mathbf{p}$ there is a riemannian isometry f which sends $\mathbf{p}$ to itself and satisfies $Df(\mathbf{p})\mathbf{x} = \mathbf{y}$.

It will also be helpful to have the following general facts about riemannian isometries.

THEOREM. *Suppose that U is as above.*

(i) If f and g are riemannian isometries of U, then so is their composite $g \circ f$, and the inverses of f and g are also riemannian isometries.

(ii) Suppose that $\mathbf{p} \in U$ and f is a riemannian isometry of U. Suppose that condition **(2′)** *above is satisfied at $\mathbf{p}$, let $\mathbf{q}$ be some other point of U, and suppose there is a riemannian isometry h of U such that $h(\mathbf{p}) = \mathbf{q}$. Then condition* **(2′)** *above is satisfied at $\mathbf{q}$.*

Sketch of proofs. The first part is a routine computation and is left to the reader as an exercise.

To prove the second part, let $\mathbf{u}$ and $\mathbf{v}$ be unit tangent vectors at $\mathbf{q}$. Then $\mathbf{x} = Df^{-1}(\mathbf{q})\mathbf{u}$ and $\mathbf{y} = Df^{-1}(\mathbf{q})\mathbf{v}$ are unit tangent vectors at $\mathbf{p}$, so by the hypothesis there is some riemannian isometry h of U which maps $\mathbf{p}$ to itself and satisfies $Dh(\mathbf{p})\mathbf{x} = \mathbf{y}$. If we take $g = f \circ h \circ f^{-1}$, then direct computation shows that g will take $\mathbf{q}$ to itself and satisfy $Dg(\mathbf{q})\mathbf{u} = \mathbf{v}$.∎

To complete the linkage of the Poincaré metric with non-Euclidean geometry, we need to check that it satisfies property **(1)** above and that it also satisfies property **(2′)** at some point of the open unit disk. By the previous theorem, these will imply that the metric also satisfies property **(2′)** at **every** point of the open unit disk.

Verification of property **(2′)** *at the origin* **0**. This turns out to be remarkably simple. Given a 2×2 orthogonal matrix A, we know that it maps the open unit disk to itself, so let f_A be this 1–1 onto mapping. It clearly has an inverse whose coordinate functions have continuous partials. Furthermore, direct calculation shows that f_A is a riemannian isometry with respect to the Poincaré metric.

Now unit vectors for the Poincaré metric at the origin are just unit vectors with respect to the standard inner product on $\mathbb{R}^2$. Given two such unit vectors, there is an orthogonal matrix taking one to the other, and since $Df_A(\mathbf{0}) = A$ it follows that property **(2′)** is satisfied at **0**.

Verification of property **(1)**. This is more difficult, so we shall first chip away at it with a sequence of reductions.

(a) *It suffices to verify the property when one of the points is the origin.* Suppose we know the property holds in this case, and let $\mathbf{p}$ and $\mathbf{q}$ be arbitrary points in the open disk. Since we are assuming the condition in the reduction, there are riemannian isometries f and g such that $f(\mathbf{0}) = \mathbf{p}$ and $g(\mathbf{0}) = \mathbf{q}$; the composite $h = g \circ f^{-1}$ then maps $\mathbf{p}$ to $\mathbf{q}$.

(b) *It suffices to verify the property when one of the points is the origin and the other is on the positive x-axis.* Suppose we know the property holds in such cases, and suppose that we have a nonzero point $\mathbf{q}$ on the unit disk. We may then write $\mathbf{q} = t\mathbf{v}$, where $\mathbf{v}$ is a unit vector and $0 < t < 1$. Since we are assuming the condition in the reduction, we have a riemannian isometry h which maps $\mathbf{0}$ to $t\mathbf{e}_1$. However, we also have an orthogonal matrix A which sends $\mathbf{e}_1$ to $\mathbf{v}$, and if f_A is the associated riemannian isometry then it will follow that $f_A \circ h$ will send $\mathbf{0}$ to $\mathbf{v}$.

(c) *Finding a riemannian isometry which sends the origin to $t\mathbf{e}_1$, where t is an arbitrary number strictly between* 0 *and* 1. This is by far the least obvious step in the whole process, and it is best done using complex numbers. Consider the following so-called *Möbius function*, which is a quotient of two linear functions defined for all complex numbers x:

$$f(z) \quad = \quad \frac{az+b}{bz+a} \qquad \text{where} \qquad a \quad = \quad \frac{1}{\sqrt{1-t^2}} \qquad \text{and} \qquad b \quad = \quad \frac{t}{\sqrt{1-t^2}}$$

This complex valued function is defined for all values of z except $-1/t$, and since $0 < t < 1$ it is defined on the open unit disk U. By construction we have $f(0) = t$ and $a^2 - b^2 = 1$.

It is probably not obvious that f sends the open unit disk U into itself. The reasons for this involve depend upon the fact that $a^2 - b^2 = 1$, and they are described in the section of the following online document titled *Disk model actions*:

`http://en.wikipedia.org/wiki/Hyperbolic_motion`

Verifying that f is a riemannian isometry requires a direct calculation of Df, which is elementary but messy. We shall omit the details.■

The preceding discussion shows that the Poincaré metric has an extensive collection of riemannian isometries, and from the viewpoint of differential geometry this is one of the key ties to non-Euclidean geometry.

The online sites listed below contain more information on the interpretation of non-Euclidean geometry using differential geometry.

`http://en.wikipedia.org/wiki/Hyperbolic_geometry`

`http://en.wikipedia.org/wiki/Upper_half_plane`

`http://mathworld.wolfram.com/PoincareHyperbolicDisk.html`

Gaussian curvature in the hyperbolic plane

Gauss' *Theorema Egregium* yields a very simple but fundamental property of the non-Euclidean plane we considered in Section III.4 of the lecture notes and the document `hyperbolic1.pdf`. The starting point is the fact that one can define the Gaussian curvature of a surface entirely in terms of its fundamental form and hence one can define Gaussian curvature for an arbitrary riemannian metric on a connected domain in $\mathbf{R}^2$. The following observation is an immediate consequence of the definition of Gaussian curvature entirely in terms of the First Fundamental Form:

PROPOSITION. *Let U be a connected domain in $\mathbf{R}^2$, and let g be a riemannian metric on U. Suppose that $f : U \to U$ is a riemannian isometry and $\mathbf{p} \in U$. Then the Gaussian curvature at $\mathbf{p}$ is equal to the Gaussian curvature at $f(\mathbf{p})$.*■

This has an immediate implication for the Poincaré metric.

COROLLARY. *The Poincaré metric has constant Gaussian curvature.*

Proof. This is true because for each pair of points $\mathbf{p}$ and $\mathbf{q}$ there is a riemannian isometry taking $\mathbf{p}$ to $\mathbf{q}$. Therefore the Gaussian curvatures at these two points are equal, and since these points are arbitrary it follows that the Gaussian curvature is the same at every point.■

It turns out that the Gaussian curvature for the Poincaré metric we have defined on the open unit disk is equal to -4; this follows from the methods used in Corollary 2.3 and Example 2.6 on pages 319 and 320 of O'Neill. Thus the open unit disk with the Poincaré metric may be viewed as an analog of the standard metrics on the Euclidean plane and sphere of radius r, which have Gaussian curvatures 0 and $1/r^2$ respectively. If we multiply the Poincaré metric by a positive constant c, then it will follow that the Gaussian curvature of the new metric is equal to $-4/c^2$ (this also follows from the references to O'Neill given above), so it the Poincaré metric and its positive multiples can have negative Gaussian curvature equal to an arbitrary negative real number.

Surfaces with constant Gaussian curvature, and their higher dimensional generalizations, play a fundamental role in differential geometry. Additional discussion of this topic appears in Section 8.6 of O'Neill.