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

We shall begin by citing the official text for the course:

> B. O'Neill. *Elementary Differential Geometry.* (Second Edition), Harcourt/Academic Press, San Diego CA, 1997, ISBN 0–112–526745–2.

This document is intended to provide a fairly complete set of notes that will reflect the content of the lectures; the approach is similar but not identical to that of O'NEILL. At various points we shall also refer to the following alternate sources. The first of these is a text at a slightly higher level, and the second is the Schaum's Outline Series review book on differential geometry, which is contains a great deal of information on the classical approach, brief outlines of the underlying theory, and many worked out examples.

> M. P. do Carmo, *Differential geometry of Curves and Surfaces*, Prentice-Hall, Saddle River NJ, 1976, ISBN 0–132–12589–7.
>
> M. Lipschultz, *Schaum's Outlines – Differential Geometry*, Schaum's/McGraw-Hill, 1969, ISBN 0–07–037985–8.

At many points we assume material covered in the preceding two 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. For the sake of completeness, a file describing the background material (with references to standard texts used in the Department's courses) is included in the course directory and can be found in the files called `background.`$*$, where $*$ is one of the extensions `dvi`, `ps`, or `pdf`.

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. 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 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). 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^{th} century, scientists and mathematicians had discovered further examples of curves that arise in various natural contexts. The development of analytic geometry and calculus, especially during the 17^{th} and 18^{th} centuries, yielded powerful new techniques for analyzing curves and 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^{th} century, largely because curves provided a means for analyzing the motion of physical objects. By the beginning of the 19^{rmth} century, the differential geometry of curves 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.

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:

`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`

I.0 : Partial differentiation

(O'Neill, § 1.1)

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 $\mathbf{R}^n$. 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 $U \subset \mathbf{R}^n$ 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).

Continuous real valued functions on open sets are defined using the same sorts of $\varepsilon - \delta$ conditions that appear in single variable calculus. 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

VECTOR LIMIT FORMULA. *Let* $\mathbf{F}$ *be a vector valued function defined on a deleted neighborhood of* $\mathbf{a}$ *with values in* $\mathbf{R}^n$, *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* $\mathbf{R}^n$, *whose* i^{th} *coordinate is equal to* 1 *and whose other coordinates are equal to zero. Then we have*

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

The previou statement about continuity of vector valued functions (continuous $\Longleftrightarrow$ all 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 $U \subset \mathbf{R}^n$, 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 $\lim_{|\mathbf{h}| \to 0} \; |\theta(\mathbf{h})| = 0$,■

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

(O'Neill, § 2.2)

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 $\mathbf{R}^3$ then their cross product or vector product is defined to be

$$\mathbf{a} \times \mathbf{b} \quad = \quad \big(a_2 b_3 - a_3 b_2, \; a_3 b_1 - a_1 b_3, \; a_1 b_2 - a_2 b_1 \big) \; .$$

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) $\mathbf{a}\times\mathbf{b} \;=\; -\mathbf{b}\times\mathbf{a}$

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

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

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

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

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

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

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

$$[\mathbf{c},\mathbf{a},\mathbf{b}] \;=\; -[\mathbf{a},\mathbf{c},\mathbf{b}] \;=\; [\mathbf{a},\mathbf{b},\mathbf{c}]\ .\blacksquare$$

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* $\mathbf{R}^3$.

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:

$$|\mathbf{a}\times\mathbf{b}|^2 \;=\; (a_2b_3 - a_3b_2)^2 + (a_3b_1 - a_1b_3)^2 + (a_1b_2 - a_2b_1)^2$$

Direct computation shows that the latter is equal to

$$(a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) \;-\; (a_1b_1 + a_2b_2 + a_3b_3)^2 \;=\; |\mathbf{a}|^2\,|\mathbf{b}|^2 \;-\; (\mathbf{a}\cdot\mathbf{b})^2$$

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

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

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

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

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 $\mathbf{R}^3$.■

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

RECOGNITION FORMULA. *If* $\mathbf{a},\mathbf{b}\in\mathbf{R}^3$ *are perpeudicular 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

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

which is the equation that we want.■

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

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

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

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

$$-|\mathbf{b}\times\mathbf{c}|^2 \;=\; -[(\mathbf{b}\times\mathbf{c}),\mathbf{b},\mathbf{c}] \;=\; [\mathbf{b},(\mathbf{b}\times\mathbf{c}),\mathbf{c}] \;=\; [\mathbf{c},\mathbf{b},(\mathbf{b}\times\mathbf{c})] \;=\;$$

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

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

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

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

$$\mathbf{a} \times \big(\mathbf{b} \times \mathbf{c}\big) \quad = \quad p\,\mathbf{b} \times \big(\mathbf{b} \times \mathbf{c}\big) \ + \ q\,\mathbf{c} \times \big(\mathbf{b} \times \mathbf{c}\big) \quad =$$

$$\Big(p\,(\mathbf{b} \cdot \mathbf{c}) + q\,(\mathbf{c} \cdot \mathbf{c})\Big)\, \mathbf{b} \ - \ \Big(p\,(\mathbf{b} \cdot \mathbf{b}) + q\,(\mathbf{b} \cdot \mathbf{c})\Big)\, \mathbf{c} \ .$$

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* $\mathbf{R}^3$ *then we have the following identity:*

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

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

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

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 it the theory of **differential forms**; a discussion of the 2- and 3-dimensional cases appears in Section II.1 of these notes.

I.2 : Parametrized curves

(O'Neill, § 1.4)

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 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

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

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. For example, the line in $\mathbf{R}^2$ 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 $V = \mathbf{R}^2$ or $\mathbf{R}^3$. 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 $\mathbf{R}^2$ as the subset of $\mathbf{R}^3$ consisting of all points whose third coordinate is zero.

Definition. A curve $\mathbf{x}$ defined on an interval J and taking values in $V = \mathbf{R}^2$ or $\mathbf{R}^3$ 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*

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

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:

(a) $\quad \mathbf{x}(t) = (t,0)$ for $t \in [0,1]$

(b) $\quad \mathbf{x}(t) = (1,t-1)$ for $t \in [1,2]$

(c) $\quad \mathbf{x}(t) = (2-t,1)$ for $t \in [2,3]$

(d) $\quad \mathbf{x}(t) = (0,1-t)$ for $t \in [3,4]$

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 an 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

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

and the Fundamental Theorem of Calculus implies that

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

Therefore if we set $R_0 = 0$ we have

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

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.

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

$$
\begin{aligned}
|R_n(h)| \quad &= \quad \left| \int_a^{a+h} \frac{(a+h-t)^n}{n!} \, \mathbf{g}^{(n+1)}(t)\, dt \right| \quad \leq \\
\operatorname{sign}(h) \quad &\cdot \quad \int_a^{a+h} \left| \frac{(a+h-t)^n}{n!} \, \mathbf{g}^{(n+1)}(t) \right| \, dt \quad \leq \\
\left(\max_{|t-a| \leq r_0} |\mathbf{g}^{(n+1)}(t)| \right) &\cdot \int_0^{|h|} \frac{u^n}{n!} \, du \quad \leq \quad M \, \frac{|h|^{n+1}}{(n+1)!}
\end{aligned}
$$

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

I.3 : Arc length and reparametrization

(O'Neill, §§ 1.4, 2.2)

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 $[0,L]$.

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 piece 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 $\mathbf{y}(u) = (u,\ \sqrt{4+u^2})$. 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

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

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

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

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 was published, a special class of such curves known as *fractal curves* has received considerable 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://academy.wolfram.agnescott.edu/ lriddle/ifs/ksnow/lsnow/htm`

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

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

I.4 : Curvature and torsion

(O'Neill, § 2.3)

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 straght 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

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

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 (M. Lipschultz, *Schaum's Outlines — Differential Geometry*, Schaum's/McGraw-Hill, 1969, ISBN 0-07-037985-8).

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

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

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 $\mathbf{R}^3$ then the plane containing them consists of all $\mathbf{x}$ such that $\mathbf{x} - \mathbf{a}$ is perpendicular to

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

which translates to the triple product equation

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

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

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

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

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

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

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

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 $\sqrt{2} = a/b$.

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

`http://mathworld.wolfram.com/Integral.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.■

Osculating planes

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

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

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

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

We then have

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

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 $\mathbf{a}$.∎

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 $\mathbf{R}^3$, 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

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

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.*

Proof. 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 $\mathbf{B}(s)$ 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 ± 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

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

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

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). 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

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

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*

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

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

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

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

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

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

(O'Neill, §§ 2.3–2.4)

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 book by O'NEILL does include a few examples, but 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 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 $\mathbf{R}^2$ or $\mathbf{R}^3$, 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:*

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

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:

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

$$
\tau(s)\,\big(\mathbf{N}(s) \times \mathbf{T}(s)\big) \;+\; \kappa\,\big(\mathbf{B}(s) \times \mathbf{N}(s)\big)
$$

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 $\mathbf{B} \times \mathbf{N} = -\mathbf{T}$. If we make these substitions into the displayed equations we obtain the second of the Frenet-Serret Formulas.■

The signifiance 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.

LOCAL UNIQUENESS FOR CURVES. *Suppose that we are given two regular smooth curves* $\mathbf{x}$ *and* $\mathbf{y}$ *defined on the same open interval containing* s_0*, where both curves are parametrized by arc length, both have continuous third derivatives and everywhere nonzero curvatures, and their curvature and torsion functions of both curves are equal. Assume further that the Frenet trihedra for both curves at* s_0 *are equal. Then* $\mathbf{y} = \mathbf{x}$ *on some open interval containing* s_0*.*

Proof. Let $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ be the standard unit vectors. We shall only consider the simplified situation where $\mathbf{x}(s_0) = \mathbf{y}(0) = \mathbf{0}$ and the Frenet trihedra for $\mathbf{x}$ and $\mathbf{y}$ at parameter value s_0 are given by $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ (one can always use a rigid motion to move the original curves into such positions, and the motion will not change the curvature or torsion of either curve — this is not really difficult to prove but it is a bit tedious and distracting).

Let $\{\,\mathbf{T}_{\mathbf{x}}(s),\, \mathbf{N}_{\mathbf{x}}(s),\, \mathbf{B}_{\mathbf{x}}(s)\}$ and $\{\,\mathbf{T}_{\mathbf{y}}(s),\, \mathbf{N}_{\mathbf{y}}(s),\, \mathbf{B}_{\mathbf{y}}(s)\}$ be the Frenet trihedra for $\mathbf{x}$ and $\mathbf{y}$ respectively, and let

$$
g(s) \;=\; |\mathbf{T}_{\mathbf{x}}(s) - \mathbf{T}_{\mathbf{y}}(s)|^2 \;+\; |\mathbf{N}_{\mathbf{x}}(s) - \mathbf{N}_{\mathbf{y}}(s)|^2 \;+\; |\mathbf{B}_{\mathbf{x}}(s) - \mathbf{B}_{\mathbf{y}}(s)|^2 \ .
$$

By the Frenet-Serret Formulas we then have that g' is equal to

$$
2\left(\Big((\mathbf{T}_{\mathbf{x}} - \mathbf{T}_{\mathbf{y}}) \cdot (\mathbf{T}'_{\mathbf{x}} - \mathbf{T}'_{\mathbf{y}}) \Big) + \Big((\mathbf{N}_{\mathbf{x}} - \mathbf{N}_{\mathbf{y}}) \cdot (\mathbf{N}'_{\mathbf{x}} - \mathbf{N}'_{\mathbf{y}}) \Big) + \Big((\mathbf{B}_{\mathbf{x}} - \mathbf{B}_{\mathbf{y}}) \cdot (\mathbf{B}'_{\mathbf{x}} - \mathbf{B}'_{\mathbf{y}}) \Big) \right) \;=
$$

$$
2\left(\Big(\kappa\ (\mathbf{T}_{\mathbf{x}} - \mathbf{T}_{\mathbf{y}}) \cdot (\mathbf{N}_{\mathbf{x}} - \mathbf{N}_{\mathbf{y}}) \Big) + \Big(\tau\ (\mathbf{B}_{\mathbf{x}} - \mathbf{B}_{\mathbf{y}}) \cdot (\mathbf{N}_{\mathbf{x}} - \mathbf{N}_{\mathbf{y}}) \Big) - \right.
$$

$$
\left. \Big(\kappa\ (\mathbf{N}_{\mathbf{x}} - \mathbf{N}_{\mathbf{y}}) \cdot (\mathbf{T}_{\mathbf{x}} - \mathbf{T}_{\mathbf{y}}) \Big) - \Big(\tau\ (\mathbf{N}_{\mathbf{x}} - \mathbf{N}_{\mathbf{y}}) \cdot (\mathbf{B}_{\mathbf{x}} - \mathbf{B}_{\mathbf{y}}) \Big) \right) .
$$

It is an elementary but clearly messy exercise in algebra to simplify the right hand side of the preceding equation, and the expression in question turns out to be zero. Therefore the function g

must be a constant, and since our assumptions imply $g(s_0) = 0$, it follows that $g(s) = 0$ for all s. The latter in turn implies that each summand

$$|\mathbf{T_x} - \mathbf{T_y}|^2 \ , \ |\mathbf{N_x} - \mathbf{N_y}|^2 \ , \ |\mathbf{B_x} - \mathbf{B_y}|^2$$

must be zero and hence that the Frenet trihedra for $\mathbf{x}$ and $\mathbf{y}$ must be the same. The first Frenet-Serret Formula then implies $\mathbf{x}' = \mathbf{y}'$, and since the two curves both go through the origin at parameter value s_0 it follows that $\mathbf{x}$ and $\mathbf{y}$ must be identical.■

There is in fact a converse to the preceding result.

FUNDAMENTAL EXISTENCE THEOREM OF LOCAL CURVE THEORY. *Given sufficiently differentiable functions κ and τ on some interval $(-c,\, c)$ such that $\kappa > 0$, there is an $h \in (0, c)$ and a sufficiently differentiable curve $\mathbf{x}$ defined on $(0, h)$ such that $\mathbf{x}(0) = \mathbf{0}$, 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*

$$\Big(\mathbf{T}(0),\ \mathbf{N}(0),\ \mathbf{B}(0)\Big) \quad = \quad \Big(\mathbf{e}_1,\ \mathbf{e}_2,\ \mathbf{e}_3\Big)$$

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

This is a consequence of the fundamental existence theorem for systems of linear differential equations. If the curve exists, then the Frenet-Serret formulas yield a system of nine first order linear differential equations for the vector valued functions $\mathbf{T}$, $\mathbf{N}$, and $\mathbf{B}$ in the Frenet trihedron

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

and if one is given κ and τ the goal is to see whether this system of first order linear differential equations can be solved for $\mathbf{T}$, $\mathbf{N}$, and $\mathbf{B}$, at least on some smaller interval $(-h,\, h)$. If one has such a solution then the curve $\mathbf{x}$ can be retrieved using the elementary formula

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

where $|s| < h$ (with the usual convention that $\int_0^s \ = \ -\int_s^0$ if $s < 0$). A proof of the existence of a solution to the system of differential equations is given on pages 309–311 in the Appendix to Chapter 4 of DO CARMO.■

The preceding two results combine to yield the **Fundamental Theorem of Local Curve Theory:**

Given κ and τ as in the statement of the Existence Theorem, 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 positive real number h_1 and 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 the restrictions of κ and τ to $(-h_1,\, h_1)$.■

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*. The following special case 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 the restriction of* $\mathbf{x}$ *to some small interval* $(s_0 - \delta,\, s_0 + \delta)$ *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 $\left(\mathbf{T}_0,\, \mathbf{N}_0,\, \mathbf{B}_0\right)$ and constant curvature $\kappa > 0$ (and also of course with vanishing torsion); in fact, the equations for an osculating circle provide an explicit construction.■

A strengthened Fundamental Theorem for plane curves

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 $\mathbf{R}^2$ 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* $\mathbf{N}$*:*

$$\begin{aligned} \mathbf{T}' &= & & k\,\mathbf{N} \\ \mathbf{N}' &= -\; k\,\mathbf{T} & & \end{aligned}$$

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

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

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

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 an* $h \in (0, c)$ *and a sufficiently differentiable curve* $\mathbf{x}$ *defined on* $(-h, h)$ *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*

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

and the curvature function is given by the restriction of κ *to* $(-h,\,h)$.■

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

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

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

$$s \quad < \quad \frac{\kappa \tau}{3!\, M} \;.■$$

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*

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

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.■

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 $\mathbf{R}^n$, 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

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

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:

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

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 site

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

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

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. The second objective 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 Chair 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

(O'Neill, § 1.5–1.6)

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 $\mathbf{R}^3$, *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.*`. 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 $\mathbf{R}^n$ 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 $\mathbf{R}^n$,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 biew as vector valued functions $\mathbf{F}$ from U to $\mathbf{R}^n$ 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:

$$x = r\cos\theta$$

$$y = r\sin\theta$$

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*

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

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 vectof 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 1-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 $\mathbf{R}^2$ or $\mathbf{R}^3$ 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 $\mathbf{R}^n$ 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

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

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 $\mathbf{R}^n$ 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.∎

II.2 : Smooth mappings

(O'Neill, §§ 1.7, 3.2)

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 $\mathbf{R}^2$ or $\mathbf{R}^3$ 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 $\mathbf{R}^n$, 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 $\mathbf{R}^n$ and that $f"U \to \mathbf{R}^m$ 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 $\mathbf{h} \in \mathbf{R}^n$ 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 $\mathbf{x}$.

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 $\mathbf{R}^n$ and $\mathbf{R}^m$ respectively, let $f : U \to V$ be a map whose coordinate functions have continuous partial derivatives at $\mathbf{x}$, and let $g : V \to \mathbf{R}^p$ 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.∎

II.3 : Inverse and implicit function theorems

(O'Neill, § 1.7)

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 texts for the Department's courses on single and multivariable calculus courses do not discuss the first result at all for functions of several variables, and only special cases of the second result are treated in one of these texts. However, 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.

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 $\mathbf{R}^n$ and $\mathbf{R}^m$ respecitvely, and let $f : U \times V \to \mathbf{R}^m$ 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 $\mathbf{R}^{n+m}$ under the standard identification of the latter with $\mathbf{R}^n \times \mathbf{R}^m$.

(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:

$$g = f^{-1} \implies g'(y) = \frac{1}{f'(g(y))}$$

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 $\mathbf{R}^n$, let $a \in U$, and let $f : U \to \mathbf{R}^n$ be a $\mathcal{C}^r$ map (where $1 \leq r \leq \infty$) 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.*`

.

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.

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

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

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. 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 $\mathbf{R}^n$, 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

(O'Neill, §§ 3.1, 3.4–3.5)

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 $\mathbf{R}^n$

Definition. Let $F : \mathbf{R}^n \to \mathbf{R}^n$ be a mapping (with no assumptions about continuity or differentiabiility). Then f is said to be an **isometry** of $\mathbf{R}^n$ if it is a 1–1 correspondence from $\mathbf{R}^n$ onto itself such that

$$|f(\mathbf{x}) - f(\mathbf{y})| = |\mathbf{x} - \mathbf{y}|$$

for all $\mathbf{x}$, $\mathbf{y} \in \mathbf{R}^n$.

Two subsets $A, B \subset \mathbf{R}^n$ are said to be **weakly congruent** if there is an isometry f of $\mathbf{R}^n$ 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 $\mathbf{R}^n$ 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

$$F\big(\mathbf{x} - \mathbf{y}\big) = \mathbf{x} - \mathbf{y} \; .$$

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 $\mathbf{R}^n$ to itself, it is meaningful to ask about its continuity or differentiability properties. The following result answers such questions simply and completely.

PROPOSITION. *Let $F : \mathbf{R}^n \to \mathbf{R}^n$ be a mapping of the form $F(\mathbf{x}) = \mathbf{b} + A(\mathbf{x})$, where $\mathbf{b} \in \mathbf{R}^n$ is some fixed vector and A is an arbitrary square matrix. Then for all $\mathbf{x} \in \mathbf{R}^n$ we have $DF(\mathbf{x}) = A$.*

COROLLARY. *Let V be open in $\mathbf{R}^m$, let $g : V \to \mathbf{R}^m$ 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 $\mathbf{R}^n$ 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.

Congruence and differentiable curves

We shall say that two continuous curves $\alpha, \beta : [a,b] \to \mathbf{R}^n$ are **congruent** if there is an isometry F of $\mathbf{R}^n$ such that $\beta = F \circ \alpha$. We are interested in the relationship between the curvatures and torsions of congruent curves.

PROPOSITION. *Let $\alpha, \beta : [a,b] \to \mathbf{R}^3$ 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 $\mathbf{R}^3$ such that $\beta = F \circ \alpha$, express F in the usual form $F(\mathbf{x}) = \mathbf{b} + A(\mathbf{x})$ where $\mathbf{b} \in \mathbf{R}^3$ and A is an orthogonal transformation, 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

$$\left(\mathbf{T}_\beta, \mathbf{N}_\beta, \mathbf{B}_\beta\right) \quad = \quad \left(A(\mathbf{T}_\alpha), A(\mathbf{N}_\alpha), A(\mathbf{B}_\alpha)\right) .$$

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

$$\mathbf{N}_\beta \quad = \quad \frac{1}{|\beta''|}\,\beta'' \quad = \quad \frac{1}{|\beta''|}\,A(\alpha'') \quad = \quad \frac{1}{|\alpha''|}\,A(\alpha'') \quad =$$

$$A\left(\frac{1}{|\alpha''|}\,\alpha''\right) \quad = \quad A(\mathbf{N}_\alpha) .$$

We must next compare the binormals; this amounts to checking whether the following cross product formula holds:

$$A(\mathbf{B}_\alpha) \quad = \quad A(\mathbf{T}_\alpha) \;\times\; A(\mathbf{N}_\alpha) \quad = \quad \mathbf{T}_\alpha \;\times\; \mathbf{N}_\beta$$

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

$$\left[A(\mathbf{T}_\alpha), A(\mathbf{N}_\alpha), A(\mathbf{B}_\alpha)\right] \quad = \quad +1 .$$

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

$$\left[A(\mathbf{T}_\alpha), A(\mathbf{N}_\alpha), A(\mathbf{B}_\alpha)\right] \quad = \quad \det A \cdot \left[\mathbf{T}_\alpha, \mathbf{N}_\alpha, \mathbf{B}_\alpha\right] \quad = \quad (+1)\cdot(+1) \quad = \quad +1$$

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 $\mathbf{R}^3$ 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 $\mathbf{R}^3$ 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

$$f(\mathbf{x}) \quad = \quad C(\mathbf{x}) \ + \ \big(\beta(s_0) - \alpha(s_0)\big)$$

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 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 $\mathbf{R}^3$. 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 $\mathbf{R}^2$. 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 $\mathbf{R}^3$ 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 $\mathbf{R}^3$ 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 a discussion of this result in the framework of these notes will be posted in the file(s) `furthertopics.*` (which can be found in the course directory).

III.1 : Mathematical descriptions of surfaces

(O'Neill, §§ 4.1, 4.8)

Some of the most basic examples of curves in $\mathbf{R}^2$ 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 $\mathbf{R}^3$ 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 $y^2 - x^2 = 1$).

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

(O'Neill, § 4.2)

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 $\mathbf{R}^2$ to $\mathbf{R}^3$ such that the 2×3 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

$$\frac{\partial \mathbf{x}}{\partial u} \quad \times \quad \frac{\partial \mathbf{x}}{\partial v}$$

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.

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*

$$\mathbf{y}(s,t) \quad = \quad \frac{\partial \mathbf{x}}{\partial u}(s,t) \ \times \ \frac{\partial \mathbf{x}}{\partial v}(s,t)$$

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*

$$D \quad = \quad \{(x,y,z) \in \mathbf{R}^3 \mid (x-s)^2 + (y-t)^2 + z^2 \quad < \quad \varepsilon^2 \ \} \ :$$

(i) The restriction of Φ *to* W *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.■

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 $\mathbf{x}(s_0, t_0)$

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 $(s,t) \in U$.■

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 $\mathbf{R}^3$ 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 $\mathbf{R}^3$ 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 $\mathbf{R}^2$ 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 $\Sigma \subset \mathbf{R}^3$ 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 $U \subset \mathbf{R}^2$ 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 $W \times \mathbf{R}$ 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 $W \times \mathbf{R}$ 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.

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 $\mathbf{R}^3$, 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 $U \subset \mathbf{R}^3$, 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

$$Ax + By + Cz = D$$

(where not all of A, B, C are zero) and spheres defined by equations of the form

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

(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 $\mathbf{R}^2$ containing (a, b) and W is an open interval in $\mathbf{R}$ 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 of the function on the left hand side 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 of the function on the left hand side 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 of the function on the left hand side 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 $\mathbf{R}^2$ defined by $y = f(x)$ and considers the set of all points $(x, y, z) \in \mathbf{R}^3$ such that $y = f(x)$. If J is the interval upon which f is defined, then this surface is the subset of $J \times \mathbf{R} \times \mathbf{R}$ 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

$$\mathbf{x}(u, v) \quad = \quad (u,\ f(u),\ v)$$

that maps $J \times \mathbf{R}$ 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

$$\Sigma(t, s) \quad = \quad (t,\ f(t),\ 0) \ + \ s \cdot (a, b, c)$$

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 $\mathbf{R}^3$ 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 $J \times \mathbf{R} \times \mathbf{R}$, 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 $\mathbf{R}^3$ 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 $\mathbf{R}^2$ 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

  $$\mathbf{X}(u,v) \quad = \quad \left(\tfrac{1}{2}(u+v),\; \tfrac{1}{2}(u-v),\; u\,v\right)\,.$$

  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 $u \in \mathbf{R}$ 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 $u \in \mathbf{R}$ 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 $u \in \mathbf{R}$ 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 $\mathbf{R}^3$ 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 M̈bius 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.*)

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 $w \in \mathbf{R}$. 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 $\mathbf{R}^3$, and let $\mathbf{p} \in \Sigma$. Define $\mathbf{K_p}$ to be the set of all vectors in $\mathbf{R}^3$ 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 $\mathbf{R}^3$.*

Proof. Let ψ be a smooth 1-1 map ψ defined on some open disk centered at $\mathbf{0}$ in $\mathbf{R}^3$ 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 $\mathbf{R}^3$ 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 $\mathbf{R}^3$ 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 $\mathbf{R}^3$ 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.

III.3 : Tangent planes

(O'Neill, § 4.3)

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 $\mathbf{R}^2$ 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 $\mathbf{R}^3$ 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.∎

III.4 : The First Fundamental Form

(O'Neill, § 4.6)

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 $\mathbf{R}^3$, 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

$$\{ (\mathbf{p}, \mathbf{q}) \in \mathbf{R}^3 \times \mathbf{R}^3 \mid \mathbf{p} \in S \ \text{ and } \ \mathbf{q} \in T_{\mathbf{p}}(S) \}$$

In some sense this consists of all the tangent planes to points in $\mathbf{R}^3$, 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 $\mathbf{R}^3$ that sends a tangent vector to its associated "free vector" (no point of application) in $\mathbf{R}^3$.

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

$$\big(x, y, z,\ X, Y, Z\big) \quad \in \quad \mathbf{R}^3 \times \mathbf{R}^3$$

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 $\mathbf{R}^3$.

Perhaps the simples 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 $U \times \mathbf{R}^2 \times \mathbf{R}^2$ 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 $\mathbf{R}^3$.

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:

$$\left(\left|\frac{\partial \mathbf{X}}{\partial u}\right|^2 \left(\frac{d\,u}{d\,t}\right)^2 \;+\; 2\left(\frac{\partial \mathbf{X}}{\partial u}\cdot\frac{\partial \mathbf{X}}{\partial v}\right)\frac{d\,u}{d\,t}\,\frac{d\,v}{d\,t} \;+\; \left|\frac{\partial \mathbf{X}}{\partial v}\right|^2 \left(\frac{d\,u}{d\,t}\right)^2\right) dt\, dt$$

Using the standard formal convention of setting

$$d\,w \quad = \quad \frac{d\,w}{d\,t}\, dt$$

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.

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

$$\mathbf{M}(u,v) \quad = \quad \begin{pmatrix} E(u,v) & F(u,v) \\ F(u,v) & G(u,v) \end{pmatrix}$$

is *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 $\mathbf{R}^2$ and the riemannian metric equal to the so-called Poincaré metric:

$$\frac{dx\,dx \;+\; dy\,dy}{(1 - x^2 - y^2)^2}$$

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

$$dt\,dt \;-\; dx\,dx \;-\; dy\,dy \;-\; dz\,dz\ .$$

III.5 : Surface area

(O'Neill, § 6.7)

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.

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 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 given in pages 2–4 of the following online document:

`http://www.math.tau.ac.il/~schuss/infi3/infi18.pdf`

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

$$\int_A |\mathbf{X}_u \times \mathbf{X}_v|\, du\, dv$$

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 no area, and this is why there is no problem using the formula even though things do not work well on the boundary.

III.6 : Curves as surface intersections

(O'Neill, § ???)

Given two distinct planes in $\mathbf{R}^3$ 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 $\mathbf{a}$ 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

$$(\frac{\sqrt{63}}{4}\cos\theta,\ \frac{1}{4},\ \frac{\sqrt{63}}{4}\sin\theta)\ .$$

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 $k(t)$.∎

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.∎

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 $f : U \to \mathbf{R}$ 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 $\mathbf{R}^3$ 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 $\mathbf{R}^3$ 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 $\mathbf{R}^3$. 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.■