I. Classical Differential Geometry of Curves

We shall begin with a few words on background material from prerequisite courses. 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`.

Differential geometry uses ideas from calculus and vector algebra to obtain geometrical information about curves and surfaces. At many points it is 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

Straight lines and circles have been central objects in geometry ever since its beginnings. During the fourth century B.C.E., Greek geometers began to study more general curves, starting with the ellipse, hyperbola and parabola. In the following centuries they discovered a large number of other curves and investigated the properties of such curves in considerable detail for a variety of reasons. The development of analytic geometry and calculus, particularly during the seventeenth and eighteenth centuries, yielded powerful 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 during that time, largely because curves provided a means for analyzing the motion of physical objects. By the beginning of the nineteenth 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. Some further results from the twentieth century will be discussed in the next unit.

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.1 : Cross products

(do Carmo, § 1–4)

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

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

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

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

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

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

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:

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

Direct computation shows that the latter is equal to

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

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 $\mathbf{R}^3$.∎

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 $\mathbf{c}$.

Since the vector $\mathbf{b} \times \big(\mathbf{b} \times \mathbf{c}\big)$ is perpendicular to $\mathbf{b} \times \mathbf{c}$ we may write it in the form

$$\mathbf{b} \times \big(\mathbf{b} \times \mathbf{c}\big) \;=\; x\,\mathbf{b} + y\,\mathbf{c}$$

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

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

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

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

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

Similarly, we also have

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

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

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

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

Since $\mathbf{b}$ and $\mathbf{c}$ are perpendicular to their cross product, the right hand side of the previous equation 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:*

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

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

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

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

I.2 : Parametrized curves

(do Carmo, § 1–2)

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*

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

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

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

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

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

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

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

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

Piecewise smooth curves

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

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

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

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

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

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

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

I.3 : Arc length and reparametrization

(do Carmo, § 1–3)

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

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

Some motivation for this definition is discused in Exercise 8 on page 10 of do Carmo. 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

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

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

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

and by the differentiation formula for inverse functions we know that

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

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

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

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

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

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

(do Carmo, §§1–5, 1–6)

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

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

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

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

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

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

$$0 \;=\; \frac{d}{ds}\Big(\mathbf{x}'(s) \cdot \mathbf{x}'(s) \Big) \;=\; 2\,\Big(\mathbf{x}'(s) \cdot \mathbf{x}''(s) \Big)$$

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

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

which translates to the triple product equation

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

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

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

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

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

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

These formulas and the basic properties of determinants imply

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

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

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

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

and hence the equation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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*

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

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

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

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

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

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*

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

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

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

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

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

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 $\tau(s) = \mathbf{B}\,'(s) \cdot \mathbf{N}(s)$.

This is not quite the same as the definition in do Carmo, so we shall show that our formulation is equivalent.

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

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

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 its torsion must be equal to zero.

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

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

which in turn implies that $\mathbf{x}\cdot\mathbf{B_0}$ is a constant. 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

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

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

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

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

(*i*) $\Big|\widehat{\mathbf{N}}\,(s) - \widehat{\mathbf{N}}\,(s_0)\Big| \leq P \cdot |\Delta s|$, 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

(do Carmo, §§1–5, 1–6, 4–Appendix)

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. A quick look at the exercises in do Carmo shows that the latter 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 should be discussed.

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 time since the appearance of do Carmo's book, spectacular advances in computer technology have provided powerful new tools for studying examples. The following book is an excellent reference for 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:*

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

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 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. Then there is a rigid motion* Φ *of 3-dimensional space such that* $\mathbf{y} = \Phi \circ \mathbf{x}$.

A rigid motion of $\mathbf{R}^2$ or $\mathbf{R}^3$ is a 1–1 and onto mapping φ such that

$$|\Phi(\mathbf{b}) \ - \ \Phi(\mathbf{a})| \ = \ |\mathbf{b} \ - \ \mathbf{a}|$$

for all vectors $\mathbf{b}$ and $\mathbf{a}$. In linear algebra it is shown that every such rigid motion has the form

$$\Phi(\mathbf{x}) \ = \ A\,\mathbf{x} + \mathbf{c}$$

where $\mathbf{c}$ is some fixed vector and A is an orthogonal matrix (*i.e.*, its rows and columns are orthonormal sets — actually, the rows are orthonormal if and only if the columns are, but we do not need this right now).

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

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

$$2\left(\Big(\kappa\,(\mathbf{T_x}-\mathbf{T_y})\cdot(\mathbf{N_x}-\mathbf{N_y})\Big) \;+\; \Big(\tau\,(\mathbf{B_x}-\mathbf{B_y})\cdot(\mathbf{N_x}-\mathbf{N_y})\Big) \;-\right.$$

$$\left.\Big(\kappa\,(\mathbf{N_x}-\mathbf{N_y})\cdot(\mathbf{T_x}-\mathbf{T_y})\Big) \;-\; \Big(\tau\,(\mathbf{N_x}-\mathbf{N_y})\cdot(\mathbf{B_x}-\mathbf{B_y})\Big)\right).$$

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*

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

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

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

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

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

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*

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

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 $\big(\mathbf{T}_0,\, \mathbf{N}_0,\, \mathbf{B}_0\,\big)$ and constant curvature $\kappa > 0$ (and also of course with vanishing torsion); in fact, the equations for an osculating circle provide an explicit construction.∎

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

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

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

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

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

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

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

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

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

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*

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

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.■

Two significant applications of the canonical form for the strong third order approximation appear on pages 28–29 of do Carmo. The proofs are elementary and contained on these pages of the text.

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* $(\mathbf{y} - \mathbf{x}(0)) > 0$*, when* $s < 0$*, and* $\mathbf{x}(s)$ *lies on the side of the osculating plane defined by the inequality* $(\mathbf{y} - \mathbf{x}(0)) <$ *point* $\mathbf{x}(s) lies on the other side of the osculating plane 0$.■

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

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

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:

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

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. Closed Curves as Boundaries

Nineteenth century techniques provided powerful means for analyzing curves quantitatively and locally. However, by the end of the nineteenth century mathematicians and users of mathematics encountered several questions of a more qualitative nature for closed curves, amy of which involve the relation between local properties and the behavior of the entire curve.. The difference between local and global properties is characterized very well on the first paragraph of page 34 of Stoker, *Differential Geometry*: A global property is one that cannot be studied by only examining the curve near each point or parameter value. In most cases, the study of global properties requires mathematical tools beyond the methods from differential calculus that dominated the first unit of the course, and frequently these come in one way or another from topology (compare the statement in the first paragraph of Section 1–7 on page 30 of do Carmo).

Let us agree that a regular smooth closed curve of class $\mathcal{C}^r$ (where $1 \leq r \leq \infty$) is a regular smooth curve $\mathbf{x} : [a, b] \to \mathbf{R}^n$, where $n = 2$ or 3, such that for each $k \leq r$ the k^{th} derivatives at a and b are equal (if $r = \infty$ the inequality should be $k < r$). Such a curve is said to be a *simple closed curve* if $\mathbf{x}(t_1) \neq \mathbf{x}(t_2)$ for $t_1 \neq t_2$ unless $t_1 = a$ and $t_2 = b$ or vice versa; in other words, it never goes through the same point twice except at the endpoints.

Here is an example of a global property that is "intuitively obvious" but definitely not easy to prove from scratch. For plane curves, examples strongly suggest that the following statement is true:

JORDAN CURVE THEOREM. *Let C be the image of a regular smooth closed curve in $\mathbf{R}^2$. Then the points of the plane not in C are contained in one of two regions A and B such that*

(i) if two points are either in A or in B, then there is a regular smooth curve joining them which lies entirely in A or B respectively,

(ii) there is a suitably large disk about the origin that contains one of the regions (the **inside** *region) and the curve itself, but there is no such disk that contains the other one (the* **inside** *region),*

(iii) if one point lies in A and the other lies in B, then every regular smooth curve joining the two points must also pass through a point of C.

The best way to understand the meaning of this result is to sit down and draw all sorts of closed curves in the plane. Each of them looks as if it has an inside region and an outside region, just like a circle. For a circle it is easy to describe the inside and outside explicitly; one is the set of all point whose distance from the center less than the radius, and the other is the set of point whose distance from the center is greater than the radius. In other relatively simple cases one can similarly describe the inside and outside regions using explicit inequalities (the reader is invited to try this with some basic examples and also to do so for *piecewise smooth* closed curves such as triangles), but in general it is hopeless to search for such descriptions. An excellent example in this connection is given online at the site

`http://ccins.camosun.bc.ca/~jbritton/fishmaze.pdf`

which also indicates the relative difficulty of determining whether two points not on the curve lie in the same or different regions. Clearly one needs additional means to attack such questions successfully.

Although the validity of the Jordan Curve Theorem was surely believed by many mathematicians and users of the subject for a long time and proofs in wide ranges of special cases were well

understood, attempts to give a mathematically rigorous proof of a general result along these lines did not begin until quite late in the nineteenth century, and the first complete proof was given by O. Veblen during the first decade of the twentieth century. In fact, he proved the result under the weaker hypothesis that the simple closed curve is merely continuous. A subsequent result of A. Schönflies yielded much stronger conclusions, including a fairly explicit description of the inside region associated to the simple closed curve. One objective of this section is to review and augment the concepts and results from multivariable calculus that are needed to formulate one version of the Schönflies Theorem. Another is to consider some applications of the Jordan Curve and Schönflies Theorems to studying the global properties of smooth regular closed curves.

The Jordan Curve Theorem is proved in many topology textbooks. Proofs for smooth curves using ideas from differential geometry are given in Section 5–7 of the text and Section 16 of the classic differential geometry text by J. J. Stoker (*Differential Geometry, etc.*). The Schönflies Theorem is somewhat more difficult to locate in textbooks; we shall discuss one proof of this result in the files `schoenflies.`$*$ that are stored in the course directory, but this proof requires material from the texts for the Department's graduate level courses in topology and complex analysis. A concise but very informative summary of the history of the Jordan Curve Theorem, the Schönflies Theorem and their analogs in higher dimensions is available online at the site

`http://math.ohio-state.edu/ fiedorow/math655/Jordan.html`

II.1 : Regions, limits and continuity

(do Carmo, 2–Appendix A, 5–Appendix)

In the discussion of the Jordan Curve Theorem we mentioned the concept of a region without saying exactly what we meant. The first order of business is to make this precise.

When working with continuous functions of a single real variable, there are only a few reasonable choices for the sets on which most functions in elementary calculus are defined. Namely, for most purposes the right sets to consider are intervals of some sort, the main choices being whether the intervals have left or right hand endpoints or are bounded or unbounded either to the left or to the right. For continuous functions of two or more variables, the situation is far more complicated even if we restrict ourselves to sets defined by reasonable inequalities (note that intervals are defined by one or more inequalities, each of which may or may not be strict). Further discussion and an example appear on pages 838–839 of *Calculus* (Seventh Edition), by Larson, Hostetler and Edwards, and Exercises 17–28 on page 846 of that book provide further examples for consideration.

In two or more dimensions, boundary sets can be extremely difficult to analyze. This contrasts sharply with the 1-dimensional situation, where the boundary generally consists of finitely many points. Most of the special plane curves described in the links at the beginning of Unit I are boundaries (or pieces from the boundaries) of fairly decent regions, but one can also go further and view most of the fractal curves from the references in Section I.3 as boundaries of regions that are not all that terrible. Mathematicians want and need to understand just how bad boundaries can be, and during the past 80 years they have developed a large array of methods for constructing regions whose boundaries are extremely wild. The subject of fractal geometry deals with some special types of irregular boundaries that are "not too wild" in an appropriate sense. If one goes from two dimensions to three, the variety of bizarre possibilities increases dramatically (the previously mentioned site on the Jordan Curve Theorem has some particularly striking pictures). For these reasons it is frequently convenient to focus on the interior (non-boundary) points of regions, and formally one does this as follows:

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. A subset U of $\mathbf{R}^n$ will be called a *connected open domain* provided

(i) for each $\mathbf{p} \in U$ there is an $r > 0$ such that the **open disk** or **ball** or **neighborhood** centered at $\mathbf{p}$ with **radius** equal to r

$$N_r(\mathbf{p}) \;=\; \left\{ \mathbf{q} \in \mathbf{R}^n \;\middle|\; |\mathbf{q} \,-\, \mathbf{p}| \;<\; r \right\}$$

is entirely contained in U,

(ii) 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 of the subsets of $\mathbf{R}^n$ that are defined by finitely many strict polynomial inequalities of the form $p_i(\mathbf{y}) \leq a_i$ satisfy (i), and either they satisfy (ii) or else they can be split into pairwise disjoint pieces such that (ii) holds on each of the pieces; one can also take the functions p_i to be standard well-behaved functions that also involve exponentials, logarithms and trigonometric functions. Of course, these sets include a vast number of central examples in multivariable calculus.

It is only necessary to make a few minor adjustments in order to work with limits and continuity on connected open domains. Once again the basic idea is that a function f is defined for all points sufficiently close to a pont $\mathbf{a}$ in some regions except perhaps at $\mathbf{a}$ itself. Since we are dealing with an open domain U this amounts to saying that there is some $r > 0$ such that the function is defined on the deleted neighborhood

$$N_r(\mathbf{a}) \ - \ \{\mathbf{a}\} \quad = \quad \left\{ \ \mathbf{x} \in \mathbf{R}^n \ \middle| \ 0 \ < \ |\mathbf{x} \ - \ \mathbf{a}| \ < \ r \right\} \ .$$

One may then say that

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

if and only if

for every $\varepsilon > 0$ there is a $\delta > 0$ such that $\mathbf{x} \neq \mathbf{a}$ and $|\mathbf{x} - \mathbf{a}| < \delta$ imply $|f(\mathbf{x}) - b| < \varepsilon$.

Continuity at $\mathbf{a}$ then means that

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

exactly as in the (opaque to the beginner) definitions from first year calculus.

Further discussion, including examples, pictures and exercises for review, may be found in Section 12.2 of *Calculus* (Seventh Edition), by Larson, Hostetler and Edwards, and pages 115–124 of *Basic Multivariable Calculus*, by Marsden, Tromba and Weinstein (full bibliographical information on these books appears in the online files `background.`$*$ mentioned at the beginning of these notes.

Limits and continuity for vector valued functions will also play an important role in this course. The quickest way to address this point is to say that a vector valued function has a limit if and only if each of its coordinate functions does, and in this case the limit of the vector valued function is the vector whose coordinates are the limits of the coordinate functions.

VECTOR LIMIT FORMULA. *Let $\mathbf{F}$ be a vector valued function defined on a deleted neighborhood of $\mathbf{a}$, 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. As usual let $\mathbf{e}_i$ denote the i^{th} unit vector in $\mathbf{R}^n$. Then we have

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

This has an important consequence:

CONTINUITY AND COORDINATE FUNCTIONS. *In the notation as above, assume that all functions are also defined at $\mathbf{a}$. Then $\mathbf{F}$ is continuous at $\mathbf{a}$ if and only if for each i, the coordinate function f_i is continuous at $\mathbf{a}$.*$\blacksquare$

Alternative approach

Here is another way of looking at limits that is an even more direct generalization of ideas from single variable calculus. Given a sequence $\{ \mathbf{x}_k$ of vectors in $\mathbf{R}^n$ we can define the limit of the sequence to be the vector whose i^{th} coordinate is the limit of the numerical sequence $\{ P_i(\mathbf{x}_k) \}$ where P_i denotes the operation of taking the i^{th} coordinate of a vector. We then have the following characterization:

LIMITS AND SEQUENCES. *Let* $\mathbf{F}$ *be a vector valued function defined on a deleted neighborhood of* $\mathbf{a}$. *Then*

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

if and only if for each sequence of vectors $\{ \mathbf{x}_k$ *in the deleted neighborhood of* $\mathbf{a}$ *whose limit is* $\mathbf{a}$ *we have*

$$\lim_{k \to \infty} \mathbf{F}(\mathbf{x}_k) \quad = \quad \mathbf{b} \ . \blacksquare$$

There is a similar statement for continuity; writing down the precise statement of the result is left to the reader.$\blacksquare$

II.2 : Smooth mappings

(do Carmo, 2–Appendix B)

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*

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

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

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

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

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

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

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

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*

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

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

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

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

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

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

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

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

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

The Jordan-Schönflies Theorem

We now have the concepts and notation necessary to state this basic result:

JORDAN-SCHÖNFLIES THEOREM *Let $\mathbf{x}$ be a regular smooth simple closed curve in the plane of class $\mathcal{C}^r$ where r is* SUFFICIENTLY LARGE, *and let Γ be the image of $\mathbf{x}$. Assume for the sake of definiteness that the domain of definition for $\mathbf{x}$ is the interval $[0,\, 2\pi]$. Then the following conclusions hold:*

(i) The complementary set $\mathbf{R}^2 - \Gamma$ consists of two connected domains, exactly one of which is bounded. Furthermore, every point of Γ is a boundary point for each of these regions.

(ii) There is a small positive number ε for which there is a $1-1$ $\mathcal{C}^r$ map $\mathbf{y}$ from the open disk $N_{1+\varepsilon}(\mathbf{0})$ in the plane into $\mathbf{R}^2$ such that $D\,\mathbf{y}$ is always invertible, the map $\mathbf{y}$ sends $N_1(\mathbf{0})$ to the bounded connected domain in $\mathbf{R}^2 - \Gamma$, and the map $\mathbf{y}$ extends $\mathbf{x}$ in the sense that $\mathbf{y}(\cos t, \sin t) = \mathbf{x}(t)$ for all $t \in [0,\, 2\pi]$.

All of the standard theorems on the global properties of plane curves involve this result to some extent. However, as noted before, the proof of this result requires concepts and methods that are far beyond the scope of this course, and the the files `schoenflies.*` explain how one can prove this result using material in standard graduate level textbooks and one other reference that is relatively accessible. For the purposes of this course it is important to understand the statement of the result, but no knowledge of the details of the proof will be needed.∎

The proof outlined in the files `schoenflies.*` actually yields a somewhat stronger conclusion.

COMPLEMENT. *In the setting of the previous result, suppose that the bounded connected domain* **Inside**(Γ) *in $\mathbf{R}^2 - \Gamma$ contains $\mathbf{0}$. Then one can choose $\mathbf{y}$ so that it is either the identity map or the reflection $S(u, v) = (u, -v)$ on some disk of radius δ, where $\delta > 0$ is chosen so that $N_{2\delta}(\mathbf{0}) \subset$* **Inside**$(\Gamma)$.∎

It is natural to ask if there is some way of determining whether $\mathbf{y}$ is given by the identity or the reflection S for points close to $\mathbf{0}$. One test for this is to compute the sign of the Jacobian of $\mathbf{y}$ at some point; since the Jacobian is nonzero everywhere (the derivative is always invertible), it turns

out that it is either positive everywhere or negative everywhere because a continuous function on a connected domain like $N_{1+\varepsilon}(\mathbf{0})$ has the Intermediate Value Property: If the function takes values a and b, then it also takes every value that lies between a and b. Since the Jacobians of the identity and S at $\mathbf{0}$ are equal to $+1$ and -1 respectively, it follows that $\mathbf{y}$ is given by the identity near $\mathbf{0}$ if the Jacobian is somewhere (hence everywhere) positive, and $\mathbf{y}$ is given by the reflection S near $\mathbf{0}$ if the Jacobian is somewhere (hence everywhere) negative.

If one wants a criterion for the behavior of $\mathbf{y}$ that is determined entirely by $\mathbf{x}$, here is one way of doing so. Choose a parameter value t^* such that the function $r(t) = |\mathbf{x}(t)|$ takes its maximum value at t^*. It is an elementary exercise to show that $\mathbf{x}'(t^*)$ is perpendicular to $\mathbf{x}(t^*)$, which will be nonzero (the only way that 0 could be the maximum value for r would be if the curve were the constant curve whose position at every parameter value is $\mathbf{0}$, but we know that the tangent vector is actually nonzero for all choices of t). If $J(u,v) = (v,-u)$ is the standard counterclockwise rotation through $\pi/2$, this means that

$$J\left(\mathbf{x}(t^*)\right) \;\; = \;\; \varepsilon\, c\, \mathbf{x}'(t^*)$$

where $\varepsilon = \pm 1$ and $c > 0$, and it turns out that the sign ε is equal to the sign of the Jacobian of $\mathbf{y}$ at $(\cos t^*, \sin t^*)$, which as we noted before is the sign of the Jacobian everywhere.

Boundary crossing formula

If we are given a simple regular smooth closed curve Γ in the plane, then trial and error suggests that we can determine whether or not two points in $\mathbf{R}^2 - \Gamma$ lie in the same connected domain by drawing a suitable regular smooth curve Λ in the plane joining the two points and counting the number of points where Λ meets Γ; if the number of crossings is even, then the two points should be in the same connected domain, while if the number of crossings is odd then they should be in different connected domains. The need for some suitability constraint is clear from very elementary considerations; in particular, if we consider a tangent line to a circle and take one point from the tangent line on each of the two rays determined by the point of tangency, then these two points lie in the exterior but the line segment joining them meets the circle in one points. In order to avoid such situations, we need to assume that the curve Λ crosses Γ *transversely* at all points of intersection; formally, this means that if $\Lambda(t_1) = \Gamma(t_2)$ then the tangent vectors $\Lambda'(t_1)$ and $\Gamma'(t_2)$ are linearly independent.

This intuitive idea can be made mathematically precise using the Jordan-Schönflies Theorem.

TRANSVERSE CROSSING CRITERION. *Let Γ be a regular smooth simple closed curve in $\mathbf{R}^2$, let $\mathbf{u}$ and $\mathbf{v}$ be points in the complement of Γ, let Λ be a regular smooth curve joining $\mathbf{u}$ to $\mathbf{v}$, and assume that the curves Γ and Λ meet in finitely many points such that each represents a transverse crossing. Then $\mathbf{u}$ and $\mathbf{v}$ lie in the same connected domain associated to $\mathbf{R}^2 - \Gamma$ if the number of crossings is even, and they lie in opposite connected domains if the number of crossings is odd.*

Note. The number of crossings is given by the number of parameter values t for Λ such that $\Lambda(t)$ lies on Γ; thus if Λ meets Γ at the same point for two separate values of t we have two distinct crossings to consider rather than only one.

Proof. The proof proceeds by induction on the number n of crossings. Suppose that $n = 0$. Then both end points of the curve must lie in the same connected domain in $\mathbf{R}^2 - \Gamma$, for if they belonged to different domains then the curve would meet Γ somewhere.

Suppose now that we know the result for curves with $(n-1)$ crossings, and suppose that Λ has n crossing points. Write these crossing points in order as follows:

$$a \;<\; t_1 \;\cdots\; <\; t_{n-1} \;<\; t_n \;<\; b$$

Choose c such that $c \in [t_{n-1}, t_n]$. Then the restriction of λ to $[a, c]$ has $(n-1)$ crossings so the induction hypothesis implies that its endpoints lie in the same connected domain if $(n-1)$ is even and in different domains if $(n-1)$ is odd. By the case $n = 0$ we also know that for every $t \in (c, t_n)$ the point $\Lambda(t)$ lies in the same connected domain as $\Lambda(c)$, and for every $t \in (t_n, b)$ the point $\Lambda(t)$ lies in the same connected domain as $\Lambda(b)$. In order to complete the inductive step, we need to show that the points of the first type lie in a different connected domain than the one containing the points of the second type.

Let Φ be the inverse function to the mapping $\mathbf{y}$ which gives a well behaved extension of Γ to the disk of radius $1 + \varepsilon$ and exists by the Jordan-Schönflies Theorem. Since $\Lambda(t_n)$ lies on Γ, by continuity there is a small constant $h > 0$ such that the restriction of Λ to $(t_n - h, t_n + h)$ lies in the image of $\mathbf{y}$; let Λ_1 be the curve in the standard disk of radius $1 + \varepsilon$ about the origin that is given by the composite of the restriction of Λ to $(t_n - h, t_n + h)$ with the inverse transformation Φ. The claim in the last sentence of the preceding paragraph will follow if we can show that $\Lambda_1(t - \frac{1}{2}\,h)$ and $\Lambda_1(t + \frac{1}{2}\,h)$ lie on opposite sides of the standard unit circle defined by the equation $x^2 + y^2 = 1$ (recall that this is the image of Γ under Φ). The latter in turn will hold if we can show that the derivative of $|\Lambda_1|^2$ at t_n is nonzero (this means that the distance from the origin is either decreasing or increasing at t_n, and we know that $|\Lambda_1(t_n)| = 1$ because Λ meets Γ at this parameter value).

Let $\mathbf{w}$ denote the tangent vector to Γ at the point $\Lambda(t_n)$. By the assumption on transverse crossings we know that $\Lambda'(t_n)$ and $\mathbf{w}$ form a basis for $\mathbf{R}^2$. Since Φ is a smooth inverse to the smooth function $\mathbf{y}$, where the derivative of the latter is always invertible, it follows that the derivative of Φ is also invertible. This implies that $\Lambda_1'(t_n)$ and the tangent vector to the unit circle at the intersection point also form a basis for $\mathbf{R}^2$. Now the tangent vector to the unit circle is a scalar multiple of $J(\,\Lambda_1(t_n)\,)$, where J is the usual counterclockwise rotation through $\pi/2$, and the linear independence of $\Lambda_1'(t_n)$ and $J(\,\Lambda_1(t_n)\,)$ is equivalent to the nonvanishing of the dot product of $\Lambda_1'(t_n)$ and $\Lambda_1(t_n)$. Since the derivative of $|\Lambda_1|^2$ is equal to

$$2\;(\Lambda_1'(t) \cdot \Lambda_1(t))$$

it follows that this derivative is nonzero at $t = t_n$, and by the preceding remarks this shows that the points of Λ at parameter values $t_n \pm \frac{1}{2}\,h$ lie in different connected domains in $\mathbf{R}^2 - \Gamma$. As noted before, this completes the proof of the inductive step.∎

II.3 : Inverse and implicit function theorems

(do Carmo, 2–Appendix B)

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 (Larson-Hostetler-Edwards and Marsden-Tromba-Weinstein) do not discuss the first result at all for functions of several variables, and only special cases of the second result are treated in Marsden-Tromba-Weinstein. 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.

BACKGROUND ON MULTIPLE INTEGRATION. Since everyone registered for this course has already taken a course that covered multiple integration and this material is really needed in order to discuss several topics in differential geometry properly, we shall assume familiarity with such material henceforth. The latter will include most material from a typical multivariable calculus course or sequence through at least some of the main theorems from vector analysis; Green's Theorem and the basic principle for recognizing gradient vector fields in two dimensions (see page 465 of the text by Marsden-Tromba-Weinstein) will definitely be needed, but the Divergence Theorem and Stokes' Theorem probably will not. Files describing the background material (with references to standard texts used in the Department's courses) is included in the course directory under the names `background2.*`.

We begin the discussion of material 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

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

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 Unit IV of the following set of notes for another course that are available online:

`http://www.math.ucr.edu/~res/math205A/prelimtext.*`

.

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`

Change of variables in multiple integrals

In multivariable calculus courses, one is interested in changes of variables arising from smooth mappings that are 1–1 and onto with Jacobians that are nonzero "almost everywhere." The standard polar, cylindrical and spherical coordinates are the most basic examples provided that one restricts the angle parameters θ and ϕ (in the spherical case) so there is no ambiguity; the Jacobian condition is reflected by the fact that this quantity is nonzero for polar and cylindrical coordinates if $r \neq 0$, and it is nonzero for spherical coordinates so long as $\rho^2 \sin \phi \neq 0$. Further discussion of this result in the general case appears on pages 333–336 of the background reference text by Marsden, Tromba and Weinstein, and on pages 995–1001 of the background reference text by Larson, Hostetler and Edwards. Exercises 37–40 on page 339 of the first reference and exercises 60–61 on page 1004 of the second are recommended as review. 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.$\blacksquare$

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 : Global properities of plane curves

(do Carmo, §§1–7, 5–7)

In these notes we shall only consider two global results explicitly, giving more details than do Carmo for one and a different perspective on another. Most global results involve integals in one way or another, and this is one reason for the choices made here. Numerous other important global results on curves (including Mukhopadhyaya's Four Vertex Theorem, the Cauchy-Crofton Theorem, Fenchel's Theorem and the Fáry-Milnor Theorem) are discused in Sections 1–7 and 5–7 of do Carmo. The following online notes also cover some of these topics with some additional details and motivation.

`http://ada.math.uga.edu/teaching/math4250/Html/Four-vertex.htm`

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

The Isoperimetric Problem

As noted in do Carmo, the *Isoperimetric Problem* is one of the oldest questions about the global properties of plane curves, going all the way back to classical Greek geometry in some form. The basic question is to find the the maximum area that can be enclosed by a simple closed curve in the plane with a given length. If one looks at various examples, it becomes apparent that maximum area is realized for circles and not for any other curves. The result itself was first demonstrated rigorously by K. Weierstrass in the nineteenth century using fairly complicated analytical methods, and subsequent research led to simplified proofs of a more geometrical nature.

ISOPERIMETRIC INEQUALITY. *Let C be a regular smooth simple closed curve in the plane of length L, and let A be the area of the union of C with its inside region (the bounded connected domain in the plane determined by the complement of C). Then $L^2 - 4\pi\, A \leq 0$, and equality holds if and only if C is a circle.*

As noted on page 35 of do Carmo, this result also holds if the smoothness hypothesis is weakened to piecewise smoothness.

Proof. The first step is a reduction to the special case where C lies in a vertical strip of the form $|x| \leq r$, it meets each of the boundary lines $x = \pm\, r$ at some points, and it lies in the region exterior to the standard circle defined by the equation $x^2 + y^2 = r^2$. To see this, follow the idea in the picture on page 34 of do Carmo and let P and Q be the maximum values for the first coordinate of $\mathbf{x}(t)$, so that the maximum and minimum first coordinates of points on the curve are P and Q respectively. It follows that the curve is contained in the vertical strip defined by the inequalities $Q \leq x \leq P$ and no strip that is properly contained in the latter. Since the simple closed curve is bounded, there is a lower bound B on its second coordinate, and one can find a circle Γ that lies in the half plane defined by the inequality $y \leq B - 1$ and is tangent to the vertical lines defined by the equations $x = Q$ and $x = P$. If $\mathbf{z}_0$ be the center of Γ and $T(\mathbf{v}) = \mathbf{v} - \mathbf{z}_0$, denotes translation by $-\mathbf{z}_0$, then the image of Γ under T is just the standard circle described above, where $2\,r = P - Q$. Since arc length and area do not change under the translation T, the theorem will be true for the original curve C if it is true for its translate $T(C)$; because of this, we shall assume for the remainder of the proof that we are working in the special case.

Choose a parametrization $\mathbf{x}(s) = (u(s), v(s))$ of C by arc length so that $\mathbf{x}(0) = \mathbf{x}(L)$ has first coordinate r and $\mathbf{x}(s_1)$ has first coordinate $-r$ for suitably chosen s_1.

We have very little *a priori* information about the connected domain that a regular simple smooth closed curve C bounds, but fortunately Green's Theorem provides several line integral formulas for the area of this region:

$$A \;=\; -\int_C y\,dx \;=\; \int_C x\,dy \;=\; \frac{1}{2}\int_C x\,dy \;-\; y\,dx$$

We want to apply these formulas both to the original curve C and to the circle Γ. However, we must be a bit careful about how we apply the formula to Γ because we need a somewhat exotic parametrization of the latter that is related to the parametrization of C. Specifically, if consider the parametrization of Γ over $[0, L]$ of the form $\mathbf{w}(s) = (X(s),\, Y(s))$ such that $X(s) = u(s)$ (*i.e.*, the same coordinate as in the parametrization of $\mathbf{x}$) and

$$Y(s) \;=\; \pm\sqrt{r^2 - (u(s))^2}$$

where the sign is positive if $s \in [0, s_1]$ and negative if $s \in [s_1, L]$; both sign choices yield the same value at s_1 because $u(s_1) = -r$ implies $Y(s_1) = \pm\, 0 = 0$. This is **not** a regular smooth parametrization, but nevertheless we can still use standard change of variables formulas from ordinary calculus to conclude that the area enclosed by Γ, which is of course equal to $\pi\, r^2$, is still given by the line integral $-\int_\Gamma y\,dx$.

The areas bounded by C and Γ are A and $\pi\, r^2$ respectively, and therefore the previously mentioned consequences of Green's theorem and the preceding paragraph show that the sums of these areas is equal to

$$A \;+\; \pi\, r^2 \;=\; \int_C x\,dy \;-\; \int_\Gamma y\,dx \;=\; \int_0^L u(s)\,v'(s)\,ds \;-\; \int_0^L Y(s)\,u'(s)\,ds\;.$$

A standard integral estimate yields the inequality

$$\int_0^L \left(u(s)\,v'(s) \;-\; Y(s)\,u'(s)\right)\,ds \;\leq\; \int_0^L \left|u(s)\,v'(s) \;-\; Y(s)\,u'(s)\right|\,ds$$

and since the integrand of the right hand expression satisfies

$$\left|u(s)\,v'(s) \;-\; Y(s)\,u'(s)\right| \;\leq\; \sqrt{u(s)^2 + Y(s)^2}\cdot\sqrt{(u'(s))^2 + (v'(s))^2}$$

by the Schwarz inequality, we may use the facts that

$$r \;=\; |\mathbf{w}(s)| \;=\; \sqrt{u(s)^2 + Y(s)^2}, \qquad 1 \;=\; |\mathbf{x}'(s)| \;=\; \sqrt{(u'(s))^2 + (v'(s))^2}$$

to conclude that the original integral is $\leq r\,L$.

We now need to use a result known as the *inequality of arithmetic and geometric means*, which states that for all nonnegative real numbers p and q we have

$$\sqrt{p\,q} \;\leq\; \frac{p+q}{2}$$

with equality if and only if $p = q$; in fact, there is an inequality of this sort if we have a set of n nonnegative real numbers for an integer $n \geq 2$, but when $n = 2$ the inequality is a straightforward consequence of the inequality

$$(\sqrt{p} - \sqrt{q})^2 \quad \geq \quad 0$$

which reduces to an equation if and only if $p = q$. If we apply this to the discussion in the preceding paragraphs, we obtain the inequalities

$$r\,L \quad \geq A \;+\; \pi\,r^2 \geq 2\,\sqrt{A\,\pi\,r^2}$$

and the inequality $L^2 \geq 4\pi\,A$ follows immediately from this.

We now need to think about what happens if equality holds; *i.e.*, we have $L^2 = 4\pi\,A$. Since we know that the inequality of the means becomes an equality if and only if the two variables are equal, it follows that we must have $A = \pi\,r^2$, which in turn implies that $L = 2\pi\,r$. Furthermore, if equality holds then the special case of the Schwarz inequality given by

$$|u(s)\,v'(s) \;-\; Y(s)\,u'(s)| \quad \leq \quad \sqrt{u(s)^2 + Y(s)^2} \cdot \sqrt{(u'(s))^2 + (v'(s))^2}$$

must be an equation for all choices of s. The resulting equation implies

$$u(s)\,u'(s) \;+\; Y(s)\,v'(s) \quad = \quad 0$$

which in turn implies that $\mathbf{x}'(s) = (u'(s),\, v'(s)\,)$ is a scalar multiple of $(-Y(s),\, u(s)\,)$. Since the first is a unit vector and the second has length r, it follows that $u(s) = \varepsilon\,r\,v'(s)$ and $Y(s) = -\varepsilon\,r\,u'(s)$, where $\varepsilon = \pm\,1$.

The next step is to prove a similar equation where the first and second coordinates are reversed. Roughly speaking, this follows by interchanging the roles of the first and second coordinates. We shall not go through the entire argument here, but instead we shall mention some key points in the argument. We can find a pair of horizontal lines $y = P'$ and $y = Q'$ such that C is tangent to each line and is contained in the horizontal strip defined by the inequalities $Q' \leq y \leq P'$. Now choose r' such that $2\,r' = P' - Q'$. As before, we next replace C by some translate $C^{\#}$ such that $C^{\#}$ is tangent to the horizontal lines $y = \pm\,r'$, it lies in the strip defined by the inequality $|y| \leq r'$, and it is reasonably far away from the circle centered at the origin with radius r'. The parametrization (U, V) for $C^{\#}$ is related to the parametrization (u, v) for C by the equations $U = u + a$, $V = v + b$ for suitable real numbers a and b. The arguments above then show that $A = \pi(r')^2$ and $L = 2\pi\,r'$, and it follows immediately that $r = r'$. Proceeding further, we obtain the equation $|V(s)| = r'\,|u'(s)|$, and since $r = r'$ this implies the desired equation $|V(s)| = r\,|u'(s)|$. It then follows that

$$\sqrt{u(s)^2 + (v(s) + b)^2} \quad = \quad \sqrt{u(s)^2 + V(s)^2} \quad = \quad r\,\sqrt{v'(s)^2 + u'(s)^2} \quad = \quad r \cdot 1 \quad = \quad r$$

and hence C is contained in the circle defined by the equation $x^2 + (y + b)^2 = r^2$.

Finally, we need to check that **every** point of this circle lies on C. Consider first the points of the form $(\pm\,r,\, b)$ We know that the first coordinates of u are equal to $\pm\,r$ when $s = 0$ and s_1. Suppose now that we are given some point on the circle whose first coordinate s^* lies in the open interval $(-r, r)$. Since the first coordinate of $\mathbf{x}$ is $+r$ when $s = 0$ or L and $-r$ when $s = s_1$, by the Intermediate Value Theorem we know that there are parameter values $w_0 \in (0, s_1)$ and $w_1 \in (s_1, L)$ such that the first coordinates of $\mathbf{x}(w_0)$ and $\mathbf{x}(w_1)$ and are equal to s^*. Since $\mathbf{x}$ is a simple closed curve, these parameter values must correspond to two separate points on the circle whose first

coordinate is s^*. Since there are exactly two points on the circle with this first coordinate, it follows that C must contain both of these points; since s^* was arbitrary, it follows that C must contain every point in the circle.∎

Curves of constant width. At one step in the proof appearing in do Carmo, it is noted that in some sense a curve with maximum area has constant width; specifically, given any direction vector in the plane, one can find two parallel lines in that direction so that the curve is entirely contained in the closed region determined by these lines and the points between them, and in fact the curve is tangent to each of these lines at some points. The only smooth curves with constant width are circles, but if one looks more generally at piecewise smooth curves there are many examples (for example, the curved triangles that the State of California puts on its highway signs that give the numbers of state highways). Here are some online references with pictures of such figures; one of the pictures is a moving demonstration of how such a curve is used to construct the so-called Wankel automobile engine:

http://mathworld.wolfram.com/CurveofConstantWidth.html

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

`http://www.keveney.com/Wankel.html`

`http://www.maa.org/mathland/mathland_10_21.html`

`http://www.cut-the-knot.org/do_you_know/cwidth.shtml`

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

`http://www.cut-the-knot.org/Curriculum/Geometry/CWStar.shtml`

`http://pauillac.inria.fr/algo/csolve/rx.pdf`

The Hopf Theorem of Turning Tangents

We shall approach this result from a slightly different perspective than the one given in do Carmo. Both proofs involve the notion of *net angular displacement* for a curve. Although the result is first stated on page 37 in Section 1–7 of do Carmo, the proof is not given until Section 5–7 because it depends upon some constructions from Section 5–6.B. This construction is important, but it takes some effort to justify it and therefore we shall give another proof of the Theorem on Turning Tangents that uses the Jordan-Schönflies Theorem and Green's Theorem.

The net angular displacement measures the extent to which a curve winds around the origin in a counterclockwise sense (hence the winding is clockwise if this quantity is negative); in order to discuss this concept, it is necessary to assume that the curve never passes through the origin. If we are given a continuous curve $\mathbf{z}$ defined on the interval $[a, b]$ and taking values in $\mathbf{R}^2 - \{\mathbf{0}\}$, then this number can be defined in two steps, first in special cases and then in general, as follows: Suppose first that the image of $\mathbf{z}$ lies in an open half plane in $\mathbf{R}^2$ defined by one of the inequalities $u_i > 0$ or $u_i < 0$ where $i = 1$ or 2 and u_i is one of the cartesian coordinates of a point in $\mathbf{R}^2$. In these cases polar coordinates define a smooth $\mathcal{C}^\infty$ maps $\mathbf{P}$ from strips of the form $(0, +\infty) \times (K, K + \pi) \subset \mathbf{R}^2$ (where K is an integral multiple of $\pi/2$ that depends upon the specific half plane) onto the given half plane; note that there are infinitely many choices of K for each half plane, and every pair of choices differs by an integral multiple of 2π. Over the strips described above, the polar coordinate map $\mathbf{P}(r, \theta)$ is 1–1 and onto, and it has a smooth inverse; in particular, the second coordinate θ of $\mathbf{P}^{-1}$ is a function such that $\tan\theta = u_2/u_1$ if the half plane is bounded by the first coordinate

axis and $\cot\theta = u_1/u_2$ if it is bounded by the second coordinate axis. In either case, it is natural to define the *net angular displacement* of $\mathbf{y}$ to be the difference between the second coordinates of $\mathbf{P}^{-1}(\,\mathbf{y}(b)\,)$ and $\mathbf{P}^{-1}(\,\mathbf{y}(a)\,)$. We have already noted that there are different possible choices for $\mathbf{P}$, and there are corresponding possibilities for its inverse, but the second coordinates for different choices of the inverse function differ by a constant (namely, an integral multiple of 2π) and therefore the difference in values does not depend upon the particular choice of $\mathbf{P}$. Similarly, if a curve lies in two distinct half planes of a given type (one with respect to the first coordinate axis and another with respect to the second), then the same considerations show that the net angular displacement does not depend upon the choice of half plane.

For a general curve $\mathbf{y}$ whose image does not lie in one of these half planes, one can define the net angular displacement as follows: A fundamental property of continuous functions called *uniform* continuity implies that we can find a $\delta > 0$ such that if $|w_1 - w_2| < \delta$ then at least one of the half planes from the previous paragraph contains both $\mathbf{y}(w_1)$ and $\mathbf{y}(w_2)$. A proof of this fact for smooth curves is described in one of the exercises for this section. Therefore, if we take a partition Π of $[a, b]$ into subintervals whose endpoints are given by consecutive points in the chain

$$a \;=\; t_0 \;<\; t_1 \;<\; \cdots \;<\; t_N \;=\; b$$

and each difference $t_i - t_{i-1}$ is less than δ, then the curves $\mathbf{y}_i$ formed by restricting $\mathbf{y}$ to the subintervals $[t_{i-1},\, t_i]$ will all have images that lie in one of the given half planes, and a natural candidate for the net angular displacement of $\mathbf{y}$ would be to take the sums of the net angular displacements for the curves $\mathbf{y}_i$.

In order to do this, we need to check two things. First, we need to show that at least one such partition exists. This is relatively easy; all we have to do is choose N such that

$$\Delta \;=\; \frac{b-a}{N} \;<\; \delta$$

and take $t_i = a + i\,\Delta$, so that each of the N subintervals has length $\Delta < \delta$. Knowing that suitable partitions exist, the second issue is to verify that this sum does not depend upon the choice of partition. Checking this second condition in general is somewhat nontrivial. A proof for arbitrary continuous curves is given in another of the exercises for this section, but in the smooth case one can deal with the problem using line integrals and the following observation,

INTEGRAL FORMULA, FIRST VERSION. *Let C be a regular smooth curve parametrized by $\mathbf{y}$, and assume that C lies in one of the four half planes described above. Then the net angular displacement of C is given by the line integral*

$$\int_C \frac{x\,dy \;-\; y\,dx}{x^2 + y^2} \;.$$

Proof. Since the polar coordinate map $\mathbf{P}$ has a smooth inverse function, we may express $\mathbf{y}$ in polar coordinates:

$$\mathbf{y}(t) \;=\; \mathbf{P}(\,r(t),\, \theta(t)\,)$$

If we substitute this into the standard formula for computing line integrals we see that the integral is equal to

$$\int_a^b \theta'(t)\,dt$$

which of course is equal to $\theta(b) - \theta(a)$.■

INTEGRAL FORMULA, SECOND VERSION. *Let C be a regular smooth curve which is parametrized by $\mathbf{y}$, and assume that C lies in $\mathbf{R}^2 - \{\mathbf{0}\}$. Then the net angular displacement of C is given by the same line integral as above.*

Proof. This follows from the integral formula for the special case in the first version along with the standard "concatenation identity"

$$\int_C \mathbf{F} \cdot d\mathbf{y} \quad = \quad \sum_{i=0}^{N} \int_{C(i)} \mathbf{F} \cdot d\mathbf{y}$$

where $C(i)$ denotes the restriction of C to the subinterval $[t_{i-1},\, t_i]$.■

For each parameter value t one can also define the net angular displacement $\alpha(t)$ up to parameter value t by applying the integral formula to the curve $\Gamma(t)$ given by the restriction of Γ to the subinterval $[a, t]$. The previous integral formulas then yield the following description of $\alpha(t)$ as an ordinary definite integral:

$$\int_a^t \frac{(x(u)\, y'(u) \;-\; y(u)\, x'(u)\,)\, du}{x(u)^2 + y(u)^2} \ .$$

We can now state and prove Hopf's Theorem:

THEOREM OF TURNING TANGENTS. *Let $\mathbf{x}$ be a regular, sufficiently differentiable, simple closed curve in $\mathbf{R}^2$ and let $\mathbf{v}(t) = \mathbf{x}'(t)$, so that the image of $\mathbf{v}$ lies in $\mathbf{R}^2 - \{\mathbf{0}\}$. Then the net angular displacement of $\mathbf{v}$ is $\pm 2\pi$.*

Some motivation and illustrations for this result appear on pages 36–37 of do Carmo.

Proof. If two curves differ by a translation, then their derivative functions are equal and therefore their net angular displacements will be equal. Therefore if $\mathbf{z}$ is a point which lies in the 'inside region" associated to the curve (more formally, the connected domain in its complement which is bounded), then the theorem will be true for $\mathbf{x}$ if and only if it is true for the translated curve $\mathbf{x}^* = \mathbf{x} - \mathbf{z}$. Since $\mathbf{0}$ belongs to the inside region for the translated curve, it follows that the proof reduces to the case where $\mathbf{0}$ is contained in the inside region associated to the curve.

As in the Jordan-Schönflies Theorem, it is convenient to assume that the curve is parametrized over $[0,\, 2\,\pi]$; this can always be achieved by a change of coordinates. The Jordan-Schönflies Theorem then implies the existence of a small positive number ε for which there is a $1 - 1$ $\mathcal{C}^r$ map

$$\mathbf{Y} : N_{1+\varepsilon}(\mathbf{0}) \to \mathbf{R}^2$$

such that $D\,\mathbf{Y}$ is always invertible, the map $\mathbf{Y}$ sends $N_1(\mathbf{0})$ to the bounded connected domain in $\mathbf{R}^2 - \Gamma$, the map $\mathbf{Y}$ extends $\mathbf{x}$ in the sense that $\mathbf{Y}(\cos t, \sin t) = \mathbf{x}(t)$ for all $t \in [0,\, 2\pi]$, and $\mathbf{Y}$ is either the identity map or the reflection $S(u, v) = (u, -v)$ on some disk of radius δ, where $\delta > 0$ is chosen so that $N_{2\delta}(\mathbf{0}) \subset \mathbf{Inside}(\Gamma)$.

The significance of $\mathbf{Y}$ is that it yields a continuous family of closed smooth curves

$$\mathbf{g}_r(t) \quad = \quad \mathbf{Y}(r\, \cos t,\, r\, \sin t)$$

with each defined on $[0,\, 2\pi]$, such that $\mathbf{g}_1 = \mathbf{x}$ and $\mathbf{g}_\delta$ is a standard parametrization of a circle of radius δ in either the counterclockwise or clockwise sense, depending on the sign of the Jacobian of $\mathbf{Y}$ at the origin. Each of these curves is regular because

$$\mathbf{g}_r'(t) \quad = \quad \frac{\partial}{\partial t}\, \mathbf{Y}(r\, \cos t,\, r\, \sin t) \quad = \quad -\, \frac{\partial \mathbf{Y}}{\partial u}\, r\, \sin t \;+\; \frac{\partial \mathbf{Y}}{\partial v}\, r\, \cos t$$

and the last expression is nonzero because (i) the invertibility of $D\mathbf{Y}$ implies that the partial derivatives of $\mathbf{Y}$ with respect to u and v are linearly independent, (ii) at least one of the coefficients $r\,\cos t$ and $r\,\sin t$ is nonzero.

For each value of r in $[\delta, 1]$ it is meaningful to discuss the net angular displacement of $\mathbf{g}'_r$. At first it may seem that this only complicates the situation by shifting attention from one curve to an infinite family of curves, but the key insight in the proof is that the exact opposite is true. As preliminary evidence for this, consider the net angular displacement for $\mathbf{g}'_\delta$. The latter is given by a relatively standard parametrization $(-\delta\,\sin t,\, \pm\,\delta\,\cos t)$, and if $T(\delta)$ is the closed curve defined by $\mathbf{g}'_\delta$, it is a standard exercise in line integrals to check that the net angular displacement for this curve is equal to

$$\int_{T(\delta)} \frac{x\,dy \;-\; y\,dx}{x^2 + y^2} \;\;=\;\; \pm\,2\pi\ .$$

The crucial idea is to show that *one obtains exactly the same value for* ALL *curves in the family* $\mathbf{g}'_r$, and this will be shown by a combination of geometric and analytic observations.

Write the smooth vector valued function $\mathbf{Y}(r\,\cos t,\, r\,\sin t)$ in coordinate form as

$$\big(\,p(r,t),\; q(r,t)\,\big)$$

and let p_2, q_2, $p_{2,2}$ and $q_{2,2}$ represent first and second partial derivatives with respect to the second variable. Then the function

$$\alpha(r,t) \;\;=\;\; \int_0^t \frac{p_2(r,u)\,q_{2,2}(r,u) \;-\; q_2(r,u)\,p_{2,2}(r,u)}{p_2^2(r,u) \;+\; q_2^2(r,u)}\,du$$

gives the net angular displacement for $\mathbf{g}'_r$ up to parameter value t, and $\alpha(r, 2\pi)$ is just the net angular displacement for $\mathbf{g}'_r$. Since everything inside the integral is continuous, we know that $\alpha(r,t)$ is a continuous function of r and t. Furthermore, we already know that $\alpha(\delta, 2\pi) = \pm\,2\pi$ and the objective is to find $\alpha(1, 2\pi)$. We shall do this by showing that the function $\alpha(r, 2\pi)$ is constant.

By definition, the angle between $\mathbf{g}'_r(0)$ and $\mathbf{g}'_r(t)$ is given by $\alpha(r,t)$ up to an integral multiple of 2π. Since $\mathbf{g}'_r(0) = \mathbf{g}'_r(2\pi)$, it follows that for each r the value of the continuous function

$$n(r) \;\;=\;\; \frac{\alpha(r, 2\pi)}{2\pi}$$

is an integer. By continuity there is a positive constant $h(r)$ so that if $s \in (r - h(r),\, r + h(r)\,)$ then $|n(s) - n(r)| < 1$, and since two distinct integers differ by at least ± 1 it follows that $n(s)$ must be equal to $n(r)$ in these cases; in other words, n is locally constant.

To see that n is globally constant, we need the following important fact about a continuous function f on a closed interval $[a, b]$: *For each value $c = f(v)$ realized by the function, there is a maximum value of $X \in [a, b]$ for which $f(X) = c$.* Applying this to n and $n(\delta)$, we obtain a maximum value M such that $n(M) = n(\delta)$. If $M = 1$ then we are done. On the other hand, if $M < 1$ then by the preceding paragraph we know that we can find some $L > M$ such that $n(L) = n(M)$, and this means that M cannot be strictly less than 1.■

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 $\mathbf{R}^2$. 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 in do Carmo there is a significant difference. In the case of 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.

One of the ultimate goals of this course 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 develops many of the basic concepts that are needed to study the differential geometry of surfaces, including the formulation and to prove a fundamental theorem for local surface theory. As in the case of curves, much of this is a generalization of material from standard multivariable calculus courses.

III.1 : Mathematical descriptions of surfaces

(do Carmo, §§2–2, 2–3)

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 else by means of the parametrization $\mathbf{r}(t) = (t,\ f(t)\,)$. 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

(do Carmo, §§2–2, 2–3)

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 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* 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

$$A\,x \;+\; B\,y \;+\; C\,z \;\;=\;\; D$$

(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

$$2\,\big(x - a,\; y - b,\; z - c\,\big)$$

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 $x^2 = a^2 > 0$. 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 $F_1(u,v)$ or the second partial $F_2(u,v)$ 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 of F is nonzero at (u,v), then the second partial of G is also nonzero at (x,y,z). If the first partial of F is nonzero at (u,v) and $x \neq 0$, then elementary calculations show that the first partial of G is also nonzero at (x,y,z), while if the first partial is nonzero and $z \neq 0$, then the third partial of G is also nonzero at (x,y,z). Since $u > a > 0$ by hypothesis we know that $x^2 + z^2 > a^2 > 0$, and therefore at least one of x and z is always zero; this proves that the gradient of G is nonzero at every point of the zero set.

RULED SURFACES. More generally, one can define another important generalization of cylindrical surfaces that also includes the cone that are **ruled** in the sense that one has parametrizations for the entire surface of the form

$$\mathbf{X}(u,v) \quad = \quad \mathbf{a}(u) + v \cdot \mathbf{b}(u)$$

where $\mathbf{a}'(u)$ is never zero and the vectors $\mathbf{a}'(u)$ and $\mathbf{b}(u)$ are always linearly independent. Here are some basic examples that are not cylindrical in the sense described above:

- A **hyperbolic paraboloid.** Consider the surface of this type defined by the equation $z = x^2 + y^2$. The right hand side factors as a product $(x-y)\,(x+y)$, so the intersection of the surface with the plane $x - y = C$ is just the line at which the planes $x - y = C$ and $z = C(x+y)$ intersect. This leads to the definition of parameters $u = x - y$ and $v = x + y$, and one can use these to parametrize the surface as

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

  Here the curves defined by holding either u or v constant are straight lines, and one can rewrite the parametrization in the form $\mathbf{y}(u) \ + \ v\,\mathbf{g}(u)$ where

  $$\mathbf{y}(u) \quad = \quad \tfrac{1}{2}\,u\,(\mathbf{e}_1 + \mathbf{e}_2)$$

  and

  $$\mathbf{g}(u) \quad = \quad \tfrac{1}{2}\,(\mathbf{e}_1 + \mathbf{e}_2) \ + \ u\,\mathbf{e}_3 \ .$$

- A **hyperboloid of one sheet.** Consider the surface of this type defined by the equation $x^2 + y^2 - z^2 = 1$. One can check directly that this surface can be parametrized using the function

  $$(\cos u,\, \sin u\,) \quad + \quad v \cdot (-\sin u,\, \cos u, 1)$$

  and that $\mathbf{a}(u) = (\cos u,\, \sin u)$ and $\mathbf{b}(u) = (-\sin u,\, \cos u, 1)$ satisfy the basic conditions described above.

- A **cone.** We shall only consider the nonsingular piece of the cone $x^2 + y^2 - z^2 = 0$ in the upper half plane where $z > 0$. In this case the parametrization is given by

  $$\mathbf{X}(u,v) \quad = \quad (v\cos u,\ v\sin u,\ v)$$

  where $u \in \mathbf{R}$ 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

$$\mathbf{X}(u,v) \;\; = \;\; (\cos u,\, \sin u, 0) \;+\; v \cdot \Big(\cos u \cos(u/2),\, \sin u \cos(u/2),\, \sin(u/2)\Big)$$

  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 Mobius strip.

`http://www.worldofescher.com/gallery/A29.html`
`http://www.mikejwilson.com/solidworks/`(*continue with next line*)
`files/mobius_II_animation.zip` (*This requires RealOne Player.*)
`http://www.physlink.com/Education/AskExperts/ae401.cfm`
`http://www.uta.edu/optics/sudduth/4d/`(*continue with next line*)
`nonorientable/moebius_strip/math/mathematics.htm`
`http://www.mapleapps.com/categories/animations/gallery/anim_pg3.shtml`
`http://www.tattva.com/vladi/director.html#6` (*Scroll down the Movie List to the last entry, which is called "Mobius strip." There are QuickTime and RealOne Player versions of this loop.*)
`http://mathworld.wolfram.com/MoebiusStrip.html` (*This is a curious animation.*)

Significant counterexamples

On the basis of our examples thus far, it is natural to ask whether the image of a parametrized surface is always a geometric surface. It turns out that the answer is negative, even if one restricts attentions to simple parametrizations that are globally 1–1. Here is one counterexample: Consider the figure 8 curve $\varphi(t) = (\sin 2t, \sin t)$ for $t \in (0,\, 2\pi)$. One then has an associated cylindrical surface with regular smooth parametrization $\Sigma(t,w) = (\sin 2t,\, \sin t,\, w)$ for $t \in (0, 2\pi)$ and $w \in \mathbf{R}$. 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.∎

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

(do Carmo, §2–4)

Special cases of tangent planes are introduced in multivariable calculus courses, particular in the case of surfaces that are graphs. 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 $bfy - \mathbf{p}$ 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

$$G(u,v,0) \;\; = \;\; (\,x(u,v),\; y(u,v),\; 0)$$

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

(do Carmo, §§2–5)

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, and 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

$$E \quad = \quad \frac{\partial \mathbf{X}}{\partial u} \cdot \frac{\partial \mathbf{X}}{\partial u} \qquad F \quad = \quad \frac{\partial \mathbf{X}}{\partial u} \cdot \frac{\partial \mathbf{X}}{\partial v} \qquad G \quad = \quad \frac{\partial \mathbf{X}}{\partial v} \cdot \frac{\partial \mathbf{X}}{\partial v}$$

This is the classical formula for the First Fundamental Form.

Abstract Riemannian metrics

In the middle of the nineteenth century G. F. B. Riemann observed that certain generalizations of the First Fundamental Form had were strongly connected to other central problems in geometry including the subject of Noneuclidean Geometry. In simplified form, his insight was to consider arbitrary expressions of the form

$$g(u,v) \;\; = \;\; E(u,v)\,du\,du \;+\; 2\,F(u,v)\,du\,dv \;+\; G(u,v)\,dv\,dv$$

where E, F and G are smooth functions on some connected domain U such that the real symmetric matrix

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

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

$$\int_a^b \; \sqrt{\big\langle \mathbf{M}\big(u(t),v(t)\,\big)\mathbf{x}'(t),\, \mathbf{x}'(t)\big\rangle}\; dt$$

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

(do Carmo, §§2–5, 2–8)

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 in Section 2–8 of do Carmo.

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:

$$\text{Area} \;=\; \int_A \sqrt{E\,G - F^2}\, du\, dv$$

The derivation of this formula from the preceding one is an elementary exercise (for example, see page 98 of do Carmo).

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

(do Carmo, §2–3)

Given two distinct planes in $\mathbf{R}3$ that have one point in common, standard axiom or theorem in 3-dimensional Euclidean geometry states that their intersection is a line. Specifically, if $\mathbf{x}$ 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 i 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 circle $x^2 + y^2 = 4$ and ellipse

$$\frac{x^2}{9} \;+\; y^2 = 1$$

about the y-axis. The intersection of these surfaces splits into two pieces, one of which consists of the point $(0, 0, 2)$ and the other of which is the circle parametrized by $(\cos\theta,\, \frac{1}{4},\, \sin\theta)$. At points of the latter the tangent planes to the two surfaces are distinct, but at $(0, 0, 2)$ 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

$$\mathbf{H}(x,y,z) \;=\; \big(\, f(x,y,z),\, g(x,y,z) \,\big)$$

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

$$F(t, u, v) \;=\; \mathbf{x}(t) \;+\; u\,\mathbf{y} \;+\; v\,\mathbf{z}\ .$$

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.∎

IV. Oriented surfaces

Given a surface Σ and a point bf p on Σ, it is meaningful to talk about the normal line to Σ at $\mathbf{p}$ which is simply the unique line that is perpendicular to the tangent plane at $\mathbf{p}$. Each such line may be viewed as having two distinct sense of direction, and an orientation is basically a way of specifying a sense of direction for every normal line to the surface. The theory of surface integrals in multivariable calculus requires the use of orientations, and Stokes' Theorem is a basic result for which orientations of surfaces are absolutely necessary; this is particularly reflected by the fact that the result does not hold for the Möbius strip. Orientations also play an important role in describing the curvature properties of a surface. Historically curvature was first defined for surfaces using orientations. Although many basic curvature properties do not depend upon orientations, the original approach to curvature using orientations provides numerous important insights that are often difficult at best to understand from other approaches.

IV.1 : Normal directions and Gauss maps

(do Carmo, §§2–6, 3–2)

If Σ is a surface and $\mathbf{p}$ is a point of Σ, then the space of tangent vectors to Σ at $\mathbf{p}$ is a 2-dimensional subspace of $\mathbf{R}^3$. The orthogonal complement of this subspace is the 1-dimensional space of normal vectors. At each point $\mathbf{p}$ there are precisely two normal vectors that have unit length.

Definition. If Σ is a surface, then an *orientation* of Σ is a continuous map $\mathbf{N} : \Sigma \to \mathbf{R}^3$ (i) such that for each $\mathbf{p} \in \Sigma$ the vector $\mathbf{N}(\mathbf{p})$ is a normal vector to $\mathbf{p}$ with unit length, (ii) for each $\mathbf{p}$ there is an open disk D centered at $\mathbf{p}$ on which $\mathbf{N}$ extends to a smooth map from D to $\mathbf{R}^3$. A surface Σ is *orientable* if one can define an orientation for Σ, and if $\mathbf{N}$ is an orientation for Σ we say that $(\Sigma, \mathbf{N})$ is an oriented surface (or surface with orientation).

Clearly orientations are not unique; in particular, if $\mathbf{N}$ is an orientation for Σ then so is $-\mathbf{N}$. Furthermore, if one considers the pair of parallel planes defined by the equation $z^2 = 1$, then one clearly has at least two choices for the orientation on each plane (namely, take $\mathbf{N} = \pm\, \mathbf{e}_3$, where the signs can be chosed independently on each of the planes). However, the following result shows that locally there are only two possible orientations for a surface.

PROPOSITION. *Let Σ be a surface, let $\mathbf{p}$ be a point of Σ, and let $\mathbf{A}$ and $\mathbf{B}$ be orientations of Σ. Then there is an open disk U containing $\mathbf{p}$ such that $\mathbf{B} = \pm\, \mathbf{A}$ on $U \cap \Sigma$.*

Proof. For each point $\mathbf{q} \in \Sigma$ we know that $\mathbf{B}(\mathbf{q}) = \pm\, \mathbf{A}(\mathbf{q})$. In particular, for each $\mathbf{q}$ this means that either $|\mathbf{B}(\mathbf{q}) \mp \mathbf{A}(\mathbf{q})| = 0$ or else $|\mathbf{B}(\mathbf{q}) \mp \mathbf{A}(\mathbf{q})| = 2$. Suppose that $\mathbf{B}(\mathbf{q}) = \mathbf{A}(\mathbf{q})$. Then by continuity we know there is some small disk D containing $\mathbf{p}$ such that $|\mathbf{B}(\mathbf{q}) \mp \mathbf{A}(\mathbf{q})| < 1$ for all $\mathbf{q}$ in $\Sigma \cap D$. Since there are only two choices for the distance $|\mathbf{B}(\mathbf{q}) \mp \mathbf{A}(\mathbf{q})|$ and one of them is greater than 1, it follows that the distance must be zero on all such points, so that $\mathbf{B}$ is equal to $\mathbf{A}$ on $\Sigma \cap D$. Similarly, if $\mathbf{B}(\mathbf{q}) = -\mathbf{A}(\mathbf{q})$, then there is an open disk V centered at $\mathbf{p}$ such that $\mathbf{B}$ is equal to $-\mathbf{A}$ on $\Sigma \cap D$.∎

There are two fundamental examples of orientable surfaces for which orientations are easy to construct.

LEVEL SURFACES. Suppose that Σ is the zero set of a smooth function f defined on a connected domain U, where as usual we assume that ∇f is always nonzero on Σ. In this case we know that the gradient of f is perpendicular to the 2-dimensional space of tangent vectors, and therefore we may define an orientation by the formula

$$\mathbf{N}(\mathbf{p}) \;=\; \frac{1}{\nabla f(\mathbf{p})} \cdot \nabla f(\mathbf{p}) \; .$$

The condition that $\mathbf{N}$ extend to a smooth disk about each point is automatically satisfied because one can use the formula to define $\mathbf{N}$ on some open disk containing $\mathbf{p}$ on which the gradient is nonzero.∎

SURFACES WITH GOOD GLOBAL PARAMETRIZATIONS. Suppose now that there is a 1–1 regular parametrization $\mathbf{X}$ for Σ that is defined on some connected domain U. Let

$$\Omega(\mathbf{p}) \;=\; \frac{\partial \mathbf{X}}{\partial u}(s,t) \;\times\; \frac{\partial \mathbf{X}}{\partial v}(s,t)$$

where $(s,t) \in U$ is the unique point such that $\mathbf{X}(s,t) = \mathbf{p}$, so that $\Omega(\mathbf{p})$ is a nonzero vector that is perpendicular to the tangent plane at $\mathbf{p}$. Then we may define an orientation by the formula

$$\mathbf{N}(\mathbf{p}) \;=\; \frac{1}{\Omega(\mathbf{p})} \cdot \Omega(\mathbf{p}) \; .$$

To verify the extension condition, let Φ be the normal thickening described in Section III.2, take

$$\Omega \;=\; \frac{\partial \Phi}{\partial u} \;\times\; \frac{\partial \Phi}{\partial v}$$

note that Ω is nonzero on an open disk containing $\mathbf{p}$, and as before take $\mathbf{N}$ to be Ω divided by its length.

A nonorientable example. The standard example is the Möbius strip; the usual way of explaining nonorientability is by studying the behavior of a unit normal vector to the surface as it travels around the central circle in the surface; as one goes around this circle once, the normal vector moves continuously from itself to minus itself. We can make this mathematically precise using the parametrization at the end of Section III.2. The central circle in the Möbius strip is just the circle in the xy-plane with parametric equations $(\cos t, \sin t, 0)$ for $t \in [0,\, 2\pi]$, and the curve with parametric equations

$$\Big(-\cos u \sin(u/2),\; -\sin u \sin(u/2),\; \cos(u/2) \Big)$$

defines a unit normal to the surface at $\mathbf{z}(t)$ for every choice of t. Direct computation shows that $\mathbf{y}(0) = (0,0,1)$ and $\mathbf{y}(2\,\pi) = (0,0,-1)$. In order to prove that the Möbius strip is not orientable, it suffices to show that this sort of thing cannot happen if one has an orientation for a surface.

PROPOSITION. *Let $\Sigma, \mathbf{N}$) be an oriented surface, let $\mathbf{z}$ be a regular smooth closed curve defined on $[0,\, 2\pi]$ and taking values in Σ, and let $\mathbf{y}$ be another smooth closed curve such that $\mathbf{y}(t)$ is a multiple of $\mathbf{N}(\mathbf{z}(t)\,)$ for all t, with $\mathbf{y}(0) = \mathbf{N}(\mathbf{z}(0)\,)$ and $\mathbf{y}(2\pi) = -\mathbf{N}(\mathbf{z}(2\pi)\,)$. Then there is a point $\alpha \in [0,\, 2\pi]$ such that $\mathbf{y}(\alpha) = \mathbf{0}$.*

Proof. The hypothesis that that $\mathbf{y}(t)$ is a multiple of $\mathbf{N}(\mathbf{z}(t)\,)$ for all t and a standard inner product formula imply that

$$\mathbf{y}(t) \;\; = \;\; \langle \mathbf{y}(t),\, \mathbf{N}(\mathbf{z}(t)\,)\,\rangle \cdot \mathbf{N}(\mathbf{z}(t)\,)$$

and therefore the coefficient

$$f(t) \;\; = \;\; \langle \mathbf{y}(t),\, \mathbf{N}(\mathbf{z}(t)\,)\,\rangle$$

is a continuous function of t. By construction we have $f(0) = 1$ and $f(2\pi) = -1$, and therefore the Intermediate Value Theorem implies the existence of some $\alpha \in (0, 2\pi)$ such that $f(\alpha) = 0$.∎

The Gauss map

The orientation map for an oriented surface is also known as the *Gauss map*. This map will play an important role in the rest of the course.

It is instructive to look at some examples. First of all, if Σ is a plane, then the normal lines at different points are all parallel to each other, and in fact if one views a plane as the set of points satisfying an equation of the form

$$A\,x \;+\; B\,y \;+\; C\,z \;\; = \;\; D$$

where $(A,\, B,\, C) \neq (0, 0, 0)$, then the natural choice for the Gauss map is the normalized gradient

$$\mathbf{N}(\mathbf{p}) \;\; = \;\; \frac{1}{\sqrt{A^2 + B^2 + C^2}} \cdot (A,\, B,\, C)$$

and accordingly the Gauss map is constant in this case. On the other hand, if one takes the sphere defined by the equation $x^2+y^2+z^2-1 = 0$, then the normalized gradient of the function at (x, y, z) is simply (x, y, z). In this case the image of the Gauss map is the entire sphere. Frequently the image of the Gauss map is somewhere between these two extremes. For example, if we consider the circular cylinder defined by the equation $x^2 + y^2 - 1 = 0$ then the normalized gradient at an arbitrary point (x, y, z) of the cylinder is $(x, y, 0)$, and the image of the Gauss map is the circle in the xy-plane defined by the equation $x^2 + y^2 - 1 = 0$. As a final example consider the parabolic cylinder $y - x^2 = 0$. The gradient at a typical point of this surface has the form $(-2x,\, 1, 0)$, and the corresponding set of unit vectors consists of all points on the unit circle in the xy-plane defined by the equation $x^2 + y^2 - 1 = 0$ **except** the point $(1, 0, 0)$.

IV.2 : The Second Fundamental Form

(do Carmo, §§3–2, 3–3)

The First Fundamental Form carries an enormous amount of information about the geometry of a surface. However, it does not completely characterize surfaces up to rigid motions in the sense of Section I.5 (*i.e*, the existence of a map Φ as in that section such that one of the surfaces is locally the image of the other under Φ). To see this, consider the plane defined by the equation $x = 1$ and the cylinder defined by $x^2 + y^2 = 1$ at the point $(1, 0, 0)$. Clearly these surfaces are not equivalent under a rigid motion. However, if one takes parametrizations near $(1, 0, 0)$ for surfaces of the forms $\mathbf{A}(u, v) = (1, u, v)$ and $\mathbf{B}(u, v) = (\cos u, \sin u, v)$ one obtains the same First Fundamental Form in terms of u and v; in each case one has $\mathbf{I} = du\,du + dv\,dv$. Physically, this reflects the fact that we can roll a flat piece of paper onto a portion of a cylinder without stretching or tearing it. Thus it is clear that we need additional data in order to characterize surfaces locally up to rigid motions. The objective of this unit is to investigate the classical description of this additional information using the Second Fundamental Form. As the name indicates, this is similar to the First Fundamental Form in some key respects, but as one might expect from the organization of these notes, its formulation requires an orientation for the surface and its definition involves the Gauss map.

Suppose that $(\Sigma, \mathbf{N})$ is an oriented surface in $\mathbf{R}^3$. We would like to define a derivative for this map $D\,\mathbf{N}$ such that for each $\mathbf{p} \in \Sigma$ we have a linear transformation $D\,\mathbf{N}(\mathbf{p})$ on the tangent space $T_{\mathbf{p}}(\Sigma)$ and in an appropriate sense $D\,\mathbf{N}(\mathbf{p})$ is a smooth function of $\mathbf{p}$. Formally, one can achieve many of these goals by taking a smooth extension $\mathbf{N}^{\#}$ of $\mathbf{N}$ on some open disk U containing $\mathbf{p}$; given a vector $\mathbf{v}$ in $T_{\mathbf{p}}(\Sigma)$ we can then provisonally define

$$[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v}) \quad = \quad \left[D\,\mathbf{N}^{\#}(\mathbf{p})\right](\mathbf{v}) \ .$$

The smoothness of this map is immediate, but it is necessary to check that the right hand side does not depend upon the choice of extension $\mathbf{N}^{\#}$ and that it sends tangent vectors at $\mathbf{p}$ to tangent vectors at $\mathbf{p}$. We shall verify these in order.

LEMMA 1. *The right hand side of the defining equation for* $[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})$ *does not depend upon the choice of extension* $\mathbf{N}^{\#}$.

Proof. Let $\mathbf{y}$ be a regular smooth curve in Σ such that $\mathbf{y}(0) = \mathbf{p}$ and $\mathbf{y}'(0) = \mathbf{v}$. Then the Chain Rule shows that the curve

$$\mathbf{z}(t) \quad = \quad \mathbf{N} \circ \mathbf{y}(t) \quad = \quad \mathbf{N}^{\#} \circ \mathbf{y}(t)$$

is also a smooth curve and $[\mathbf{N} \circ \mathbf{y}]'(0)$ is equal to the right hand side of the defining equation for $[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})$. Since $\mathbf{z}'(0) = [\mathbf{N} \circ \mathbf{y}]'(0)$ does not depend upon the choice of extension, this proves that the definition for $[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})$ also does not depend upon the choice of extension.∎

LEMMA 2. *If* $\mathbf{v} \in T_{\mathbf{p}}(\Sigma)$ *then* $[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})$ *also lies in* $T_{\mathbf{p}}(\Sigma)$.

Proof. We shall use the notation of the preceding lemma, so that

$$[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v}) \quad = \quad \mathbf{z}'(0)$$

where $\mathbf{z} = \mathbf{N} \circ \mathbf{y}$. It suffices to show that $\mathbf{z}'(0)$ is perpendicular to $\mathbf{N}(\mathbf{p})$. For each $\mathbf{q} \in \Sigma$ the normal vector $\mathbf{N}(\mathbf{q})$ has unit length by construction, and therefore we know that $|\mathbf{z}'(t)|^2 \equiv 1$. If

we differentiate this and apply the Leibniz Rule for dot products of vector valued functions, we see that

$$0 \;=\; \frac{d}{dt}\,|\mathbf{z}'(t)|^2 \;=\; 2\,\langle \mathbf{z}(t),\, \mathbf{z}'(t)\,\rangle$$

and if we evaluate the inner product on the right hand side at $t = 0$ we see that it is equal to

$$\langle \mathbf{N}(\mathbf{p}),\, [D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})\,\rangle$$

and therefore we conclude that this inner product vanishes; *i.e.*, the vector $[D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{v})$ is perpendicular to $\mathbf{N}(\mathbf{p})$.∎

If we are given a local parametrization of Σ near $\mathbf{p}$ by some regular smooth parametrization $\mathbf{X}$, defined on some connected domain U, then we can use the local formula for $\mathbf{N}$ to describe the map

$$\left[D\left(\mathbf{N}^{\circ}\mathbf{X}\right)\right]\;(u,v)$$

as a linear transformation on $\mathbf{R}^2$ using the identity

$$\mathbf{N}^{\circ}\mathbf{X} \;=\; \frac{1}{(\text{length})}\cdot\left(\frac{\partial \mathbf{X}}{\partial u}\right)\;\times\left(\;\frac{\partial \mathbf{X}}{\partial v}\right)$$

but we shall find a better way to compute this map in terms of local coordinates.

In textbooks and elsewhere the map $D\,\mathbf{N}(\mathbf{p})$ is often called the *Weingarten map* or the *shape operator* for the oriented surface $(\Sigma,\, \mathbf{N})$.

Definition. Let $\mathbf{T}^{(2)}(\Sigma)$ be the set of all ordered pairs of points $(\mathbf{v}_1,\, \mathbf{v}_2)$ in $\mathbf{T}(\Sigma) \times \mathbf{T}(\Sigma)$ such that $\tau_\Sigma(\mathbf{v}_1) = \tau_\Sigma(\mathbf{v}_2)$, and let $\mathbf{x}_i$ be the second coordinate of $\mathbf{v}_i$. The *Second Fundamental Form* of Σ is the map $\mathbf{II}_\Sigma$ on $\mathbf{T}^{(2)}(\Sigma)$ sending $(\mathbf{v}_1,\, \mathbf{v}_2)$ to

$$-\,\langle [D\,\mathbf{N}(\mathbf{p})]\,(\mathbf{x}_1),\, \mathbf{x}_2\rangle$$

(note the sign!!!) where as usual $\langle \ldots,\, \ldots\rangle$ denotes the usual inner product of two vectors in $\mathbf{R}^3$.

At this point it is helpful to understand what this means for the basic examples we discussed in the previous section. If Σ is a plane and $\mathbf{N}$ is some orientation, then $\mathbf{N}$ is parallel to the normal direction for the plane at all points and therefore $\mathbf{N}$ is at least locally constant by the observations in Section IV.1. In fact, one can take the extension $\mathbf{N}$ also to be constant, and it follows that in this case $D\,\mathbf{N}(\mathbf{p}) = 0$ for all points $\mathbf{p}$ on the plane, and therefore the Second Fundamental Form is also zero at every point of the plane.. On the other hand, if Σ is the sphere and $\mathbf{N}$ is the standard outward pointing orientation then $\mathbf{N}(\mathbf{p}) = \mathbf{p}$ and therefore $D\,\mathbf{N}(\mathbf{p}) = I$ for all points $\mathbf{p}$ on the sphere; in this case the Second Fundamental Form is the negative of the First Fundamental Form. Of course, if we would replace $\mathbf{N}$ by its negative, then the Second and First Fundamental Forms would be equal. Finally, suppose that Σ is the cylinder defined by $x^2 + y^2 = 1$. In this case $\mathbf{N}(x, y, z) = (x, y, 0)$ and the tangent space at (x, y, z) has an orthonormal basis given by $(-y, x, 0)$, which is the tangent vector to the circle $\mathbf{a}(\theta) = (\cos\theta,\, \sin\theta,\, z)$ at (x, y, z), and (0.0.1), which is the tangent vector to the vertical line $\mathbf{b}(t) = (x, y, t)$ at (x, y, z). Direct computation shows that these two vectors are eigenvectors for $D\,\mathbf{N}(x, y, z)$ and the associated eigenvalues are 1 and 0 respectively. It follows that the Second Fundamental Form is given by

$$-\,\langle [D\,\mathbf{N}(\mathbf{p})]\,(p\,\mathbf{a}' \;+\; q\,\mathbf{b}'),\, r\,\mathbf{a}' \;+\; s\,\mathbf{b}'\,\rangle$$

and by the observations in the preceding sentences this is equal to $-pr$. In particular, the Second Fundamental Form vanishes for some but not all pairs of vectors. At the beginning of this section we mentioned that the First Fundamental Forms for the plane and the cylinder were the same, but we have now seen that their Second Fundamental Forms are different.

We would also like to understand the behavior of the Second Fundamental Form for the hyperboloid of one sheet given by the equation $x^2 - y^2 - z^2 - 1 = 0$, and in order to do this we need to follow through on our earlier comment about developing a way of computing $D\,\mathbf{N}$ and the Second Fundamental Form in a efficiently using a regular smooth parametrization $\mathbf{X}$. The first point is to observe that the Second Fundamental Form is completely determined by its values for pairs of tangent vectors such that each is either $\mathbf{X}_u$ and $\mathbf{X}_v$ (this includes cases where both vectors in the pair are the same and where the two vectors are different). This is true because

$$\mathbf{II}\,(p\,\mathbf{X}_u \;+\; q\,\mathbf{X}_v),\, r\,\mathbf{X}_u \;+\; s\,\mathbf{X}_v\,) \;\;=$$

$$p\,r\,\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_u) \;+\; p\,s\,\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_v) \;+\; q\,r\,\mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_u) \;+\; q\,s\,\mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_v)\;.$$

We then have the following important formulas:

BASIC VALUES FOR SECOND FUNDAMENTAL FORMS. *Suppose we are given a regularly parametrized smooth surface Σ with parametrization $\mathbf{X}$, and assume that the normal vector function is given by $\mathbf{N}$. Then the following identities hold for the second fundamental form:*

[1] $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_u,) = \langle \mathbf{N},\,\mathbf{X}_{u,u}\,\rangle$

[2] $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_v,) = \langle \mathbf{N},\,\mathbf{X}_{u,v}\,\rangle = \langle \mathbf{N},\,\mathbf{X}_{v,u}\,\rangle = \mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_u,)$

[3] $\mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_v,) = \langle \mathbf{N},\,\mathbf{X}_{v,v}\,\rangle$

Proof. All the derivations are of a similar nature, so we start with the first one. Since $\mathbf{N}$ is perpendicular to the tangent plane at every point we know that

$$\langle \mathbf{N},\,\mathbf{X}_u\,\rangle \;\;=\;\; 0\;.$$

Next, observe that $D\,\mathbf{N}(\mathbf{p})\,\mathbf{X}_u$ is merely the partial derivative $D_u\,(\mathbf{N}\circ\mathbf{X})$, and we shall stretch our conventions to write this as $D_u\,\mathbf{N}$. Taking partial derivatives of the equation above with respect to u and applying the Leibniz Rule for dot products of vector valued functions, we see that

$$0 \;\;=\;\; \langle D_u\,\mathbf{N},\,\mathbf{X}_u\,\rangle \;+\; \langle \mathbf{N},\,\mathbf{X}_{u,u}\,\rangle$$

and since the first summand on the right hand side is the negative of $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_u,)$ it follows that the latter is equal to $\langle \mathbf{N},\,\mathbf{X}_{u,u}\,\rangle$ as required. The derivation of the third identity is nearly identical, the only difference being that u is replaced by v in each equation.

Similarly, if we take partial derivatives of both sides of the equation

$$\langle \mathbf{N},\,\mathbf{X}_u\,\rangle \;\;=\;\; 0$$

with respect to v we conclude that $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_v) = \langle \mathbf{N},\,\mathbf{X}_{u,v}\,\rangle$. Furthermore, if we interchange the roles of u and v in this argument we also see that $\mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_u) = \langle \mathbf{N},\,\mathbf{X}_{v,u}\,\rangle$. Under the assumption that the regular surface parametrization has continuous second partial derivatives, we know that $\mathbf{X}_{u,v} = \mathbf{X}_{v,u}$ and using this we see that $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_v) = \mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_u)$. This completes the derivation.∎

Notational conventions. In the literature and textbooks the quantities $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_u\,)$, $\mathbf{II}\,(\mathbf{X}_u,\,\mathbf{X}_v\,) = \mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_u\,)$ and $\mathbf{II}\,(\mathbf{X}_v,\,\mathbf{X}_v\,)$ are often denoted by e, f and g respectively or by L, M and N respectively; in these notes we shall use the first notation in order to avoid confusion between N and $\mathbf{N}$. If one writes a typical tangent vector in the form $\mathbf{X_u}\,du + \mathbf{X_v}\,dv$ where du and dv are viewed as scalars, then this yields the classical expression for the Second Fundamental Form:

$$\mathbf{II}\,(\mathbf{X}_u\,du + \mathbf{X}_v\,dv,\,\mathbf{X}_u\,du + \mathbf{X}_v\,dv\,) \;\; = \;\;$$

$$e(u,v)\,du\,du \;\; + \;\; 2\,f\,(u,v)\,du\,dv \;\; + \;\; g(u,v)\,dv\,dv$$

Example. We shall apply all this to describe the Second Fundamental Form for the hyperboloid of one sheet described above. If we let $h(x,y,z) = x^2 + y^2 - z^2 - 1$ then $\nabla h(x,y,z) = 2\,(x,y,-z)$ and therefore the unit normal is given by

$$\mathbf{N}(x,y,z) = \frac{1}{\sqrt{x^2+y^2+z^2}} \cdot (x,y,-z)\;.$$

This can be simplified slightly by noting that $x^2 + y^2 + z^2 = 1 + 2\,z^2$ for points on the surface, but it will also be useful for us to let $\Omega(x,y,z) = (x,y,-z)$, so that $\Omega(x,y,z)$ is a positive multiple of the unit normal $\mathbf{N}$ described above.

We shall use the following standard parametrization for the hyperboloid of one sheet:

$$\mathbf{X}(u,v) \;\; = \;\; \big(\cos u\,\cosh v,\,\sin u\,\cosh v,\,\sinh v\big)$$

One can then describe the normal vector Ω by the following formula:

$$\Omega(u,v) \;\; = \;\; \big(\cos u\,\cosh v,\,\sin u\,\cosh v,\,-\sinh v\big)$$

One can compute the second partial derivatives of $\mathbf{X}$ in a very direct manner, and if one takes inner products with Ω one obtains the following results:

$$\langle\,\Omega,\,\mathbf{X}_{u,u}\,\rangle = -\cosh^2 v$$

$$\langle\,\Omega,\,\mathbf{X}_{u,v}\,\rangle = \langle\,\Omega,\,\mathbf{X}_{v,u}\,\rangle = 0$$

$$\langle\,\Omega,\,\mathbf{X}_{v,v}\,\rangle = 1$$

Therefore it follows that the coefficients $e(u,v)$ are always negative, the coefficients $f(u,v)$ are always zero and the coefficients $g(u,v)$ are always positive. Thus the Second Fundamental Form in this case looks quite different from those in the cases previously described.

Since many surfaces are expressed as graphs of functions of two variables, we shall describe the First and Second Fundamental Forms for a surface defined as the graph of a smooth function $z = h(x,y)$. In order to state these formulas concisely we shall let $\alpha(x,y) = \sqrt{1 + h_x^2 + h_y^2}$.

$$E = 1 + h_x^2$$

$$F = h_x\,h_y$$

$$G = 1 + h_y^2$$

$$e = h_{x,x}/\alpha$$

$$f = h_{x,y}/\alpha$$

$$g = h_{y,y}/\alpha$$

All of these formulas may be verified directly using the identities established above and the fact that the regular smooth parametrization of the surface is given by $\mathbf{X}(x,y) = (x,y,\,h(x,y)\,)$.

IV.3 : Quadratic forms and adjoint transformations

(do Carmo, §3–Appendix)

The First and Second Fundamental Forms are examples of quadratic forms on a real vector space with an inner product. It is particularly useful to study some aspects of the Second Fundamental Form using a few basic algebraic facts about such quadratic forms, so we shall summarize what is needed here. For our purposes it will suffice to restrict our attention to quadratic forms on 2-dimensional real inner product spaces.

The following result is an easy algebraic exercise:

PROPOSITION. *Let V be a 2-dimensional real inner product space with basis vectors $\mathbf{x}$ and $\mathbf{y}$, and let $T : V \to V$ be a linear transformation from V to itself. If $\langle T(\mathbf{x}),\, \mathbf{y}\rangle = \langle \mathbf{x},\, T(\mathbf{y})\,\rangle$, then $\langle T(\mathbf{w}),\, \mathbf{z}\rangle = \langle \mathbf{w},\, T(\mathbf{z})\,\rangle$ for all $\mathbf{w},\, \mathbf{z} \in V$.*

Proof. Express $\mathbf{w}$ and $\mathbf{z}$ as linear combinations of $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{w} \;=\; p\,\mathbf{x} \,+\, q\,\mathbf{y} \qquad \mathbf{z} \;=\; r\,\mathbf{x} \,+\, s\,\mathbf{y}$$

We then have

$$\langle T(\mathbf{w}),\, \mathbf{z}\rangle \;=\; pr\,\langle T(\mathbf{x}),\, \mathbf{x}\rangle \,+\, qr\,\langle T(\mathbf{y}),\, \mathbf{x}\rangle \,+\, ps\,\langle T(\mathbf{x}),\, \mathbf{y}\rangle \,+\, qs\,\langle T(\mathbf{w}),\, \mathbf{z}\rangle$$

and similarly we have

$$\langle \mathbf{w},\, T(\mathbf{z})\,\rangle \;=\; pr\langle \mathbf{x},\, T(\mathbf{x})\,\rangle \;=\; qr\,\langle \mathbf{x},\, T(\mathbf{y})\,\rangle \;=\; ps\,\langle \mathbf{y},\, T(\mathbf{x})\,\rangle \;=\; qs\,\langle \mathbf{y},\, T(\mathbf{y})\,\rangle\ .$$

We always have

$$\langle T(\mathbf{x}),\, \mathbf{x}\rangle = \langle \mathbf{x},\, T(\mathbf{x})\,\rangle$$

and similarly if $\mathbf{y}$ replaces $\mathbf{x}$, so the hypothesis $\langle T(\mathbf{x}),\, \mathbf{y}\rangle = \langle \mathbf{x},\, T(\mathbf{y})\,\rangle$ combines with these to show that $\langle T(\mathbf{w}),\, \mathbf{z}\rangle = \langle \mathbf{w},\, T(\mathbf{z})\,\rangle$ for all $\mathbf{w},\, \mathbf{z} \in V$.∎

Linear transformations satisfying the conclusion of the preceding result are said to be *self-adjoint*. If we are given an orthonormal basis $\mathbf{u}$ and $\mathbf{v}$ for our inner product space V and we construct the 2×2 matrix representing T with this orthonormal basis

$$T(\mathbf{u}) \;=\; a\,\mathbf{u} \,+\, b\,\mathbf{v} \qquad T(\mathbf{v}) \;=\; c\,\mathbf{u} \,+\, d\,\mathbf{v}$$

then T is self adjoint if and only if

$$c = \langle \mathbf{u},\, T(\mathbf{v})\,\rangle \;=\; \langle T(\mathbf{u}),\, \mathbf{v}\,\rangle \;=\; b$$

or in other words the matrix representing T is *symmetric*.

Every real symmetric matrix has an orthonormal basis of eigenvectors; this is a standard result on matrices, and in the 2×2 case one can see this very easily by computing the characteristic polynomial and noting that it has real roots. Other basic results in linear algebra imply a corresonding result of this sort for self-adjoint linear transformations.

Let V and T be as above. In the next section we shall be interested in finding the maximum and minimum values of the quotient

$$k(\mathbf{x}) \quad = \quad \frac{\langle T(\mathbf{x}),\, \mathbf{x}\,\rangle}{\langle \mathbf{x},\, \mathbf{x}\,\rangle}$$

where $\mathbf{x}$ ranges over all nonzero vectors in V. Let $\mathbf{u}$ and $\mathbf{v}$ be an orthonormal basis of eigenvectors for T, and let α and β be the eigenvalues associated to $\mathbf{u}$ and $\mathbf{v}$ respectively. One of these eigenvalues is greater than or equal to the other, and we shall assume that we have labeled everything so that $\alpha \leq \beta$.

RAYLEIGH'S PRINCIPLE. *The maximum and minimum values of the above expression are the eigenvalues β and α, and these values are attained and the eigenvectors $\mathbf{v}$ and $\mathbf{u}$ respectively.*

Proof. If $\mathbf{x}$ is an arbitrary nonzero vector in V we may write

$$\mathbf{x} \quad = \quad r\,\cos\theta\,\mathbf{x} \;+\; r\,\sin\theta\,\mathbf{y}$$

for some $r > 0$ and $\theta \in \mathbf{R}$. It follows that

$$k(\mathbf{x}) = \alpha\,\cos^2\theta + \beta\,\sin^2\theta\ .$$

Since $\alpha \leq \beta$ it follows that

$$\alpha \quad = \quad \alpha\,\cos^2\theta + \alpha\,\sin^2\theta \quad \leq \quad \alpha\,\cos^2\theta + \beta\,\sin^2\theta \quad \leq \beta\,\cos^2\theta + \beta\,\sin^2\theta \quad = \quad \beta$$

and it also follows that $k(\mathbf{u}) = \alpha$ while $k(\mathbf{v}) = \beta$.∎

Trace and determinant formulas

Recall that the trace of a square matrix is equal to the sum of its diagonal entries, and if a matrix is diagonalizable then the trace is equal to the weighted sum of the eigenvalues

$$\sum_{\lambda}\, n(\lambda)\,\lambda$$

(one way to see this is by means of the characteristic polynomial — both numbers are $(-1)^{n-1}$ times the coefficient of t^{n-1} if the matrix in question is $n \times n$). One can then define the trace of a diagonalizable linear transformation on a finite-dimensional vector space by means of the corresponding weighted sum of eigenvalues. This is entirely analogous to the situation for the determinant. For diagonalizable matrices the latter is equal to the weighted product of eigenvalues

$$\sum_{\lambda}\, n(\lambda)\,\lambda$$

and one can define the determinant of a diagonalizable linear transformation on a finite dimensional vector space using this formula.

We shall need information about the following purely algebraic question:

Problem. Suppose that we are given a 2-dimensional real inner product space V, a basis $\mathbf{z}_1$ and $\mathbf{z}_2$ for V, and a self-adjoint linear transformation $T : V \to V$. Suppose that

$$A \;=\; \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

is the matrix whose entries are defined by the formulas

$$T(\mathbf{z}_1) \;=\; a\,\mathbf{x}_1 \,+\, b\,\mathbf{z}_2 \quad \text{and} \quad T(\mathbf{z}_1) \;=\; a\,\mathbf{x}_1 \,+\, b\,\mathbf{z}_2 \;.$$

Express the entries of A in terms of the inner products $\langle \mathbf{z}_i,\, \mathbf{z}_j \,\rangle$ and $\langle T(\mathbf{z}_i),\, \mathbf{z}_j \,\rangle$ where $1 \le i, j \le 2$.

Motivated by our terminology for the First and Second Fundamental Forms, we shall denote the various inner products as indicated in the matrices below:

$$\begin{pmatrix} \langle \mathbf{z}_1,\, \mathbf{z}_1 \,\rangle & \langle \mathbf{z}_1,\, \mathbf{z}_2 \,\rangle \\ \langle \mathbf{z}_2,\, \mathbf{z}_1 \,\rangle & \langle \mathbf{z}_2,\, \mathbf{z}_2 \,\rangle \end{pmatrix} \;=\; \begin{pmatrix} E & F \\ F & G \end{pmatrix} \qquad \begin{pmatrix} \langle T(\mathbf{z}_1),\, \mathbf{z}_1 \,\rangle & \langle T(\mathbf{z}_1),\, \mathbf{z}_2 \,\rangle \\ \langle T(\mathbf{z}_2),\, \mathbf{z}_1 \,\rangle & \langle T(\mathbf{z}_2),\, \mathbf{z}_2 \,\rangle \end{pmatrix} \;=\; \begin{pmatrix} e & f \\ f & g \end{pmatrix}$$

Direct calculuation then yields the following equations:

$$\begin{aligned}
e \;&=\; \langle T(\mathbf{z}_1),\, \mathbf{z}_1 \,\rangle \;=\; \langle T(a\,\mathbf{z}_1 \,+\, b\,\mathbf{z}_2,\, \mathbf{z}_1 \,\rangle \;=\; a\,E \,+\, b\,F \\
f \;&=\; \langle T(\mathbf{z}_2),\, \mathbf{z}_1 \,\rangle \;=\; \langle T(c\,\mathbf{z}_1 \,+\, d\,\mathbf{z}_2,\, \mathbf{z}_1 \,\rangle \;=\; c\,E \,+\, d\,F \\
f \;&=\; \langle T(\mathbf{z}_1),\, \mathbf{z}_2 \,\rangle \;=\; \langle T(a\,\mathbf{z}_1 \,+\, b\,\mathbf{z}_2,\, \mathbf{z}_2 \,\rangle \;=\; a\,F \,+\, b\,G \\
g \;&=\; \langle T(\mathbf{z}_2),\, \mathbf{z}_2 \,\rangle \;=\; \langle T(c\,\mathbf{z}_1 \,+\, d\,\mathbf{z}_2,\, \mathbf{z}_1 \,\rangle \;=\; c\,F \,+\, d\,G
\end{aligned}$$

These equations are equivalent to the following matrix equation:

$$\begin{pmatrix} e & f \\ f & g \end{pmatrix} \;=\; \begin{pmatrix} a & c \\ b & d \end{pmatrix} \cdot \begin{pmatrix} E & F \\ F & G \end{pmatrix}$$

If we now assume that the matrix with entries E, F and G is invertible (as it is in the case of the First Fundamental Form), then one can solve for A and obtain descriptions of its entries in terms of the entries of the other two matrices. These lead directly to the identities we want:

TRACE AND DETERMINANT FORMULAS. *The determinant of A is equal to*

$$\frac{e\,g - f^2}{E\,G - F^2}$$

and the trace of A is equal to

$$\frac{e\,G - 2\,f\,F + g\,E}{E\,G - F^2} \;.$$

Derivation. The first of these follows from the matrix equation and the fact that $\det(B_1 B_2) = \det B_1 \cdot \det B_2$. For the second formula we need to compute

$$\begin{pmatrix} e & f \\ f & g \end{pmatrix} \cdot \begin{pmatrix} E & F \\ F & G \end{pmatrix}^{-1}$$

and take the sum of its diagonal entries. By Cramer's Rule the inverse is given by

$$\frac{1}{E\,G - F^2} \cdot \begin{pmatrix} G & -F \\ -F & E \end{pmatrix}$$

and if substitute this into the preceding formula, compute the product, and add the diagonal entries then we obtain the expression for the trace in the formula.∎

IV.4 : Normal, Gaussian and mean curvature

(do Carmo, §§3–2, 3–3)

One approach to studying the curvature properties of surfaces is to consider the curvature properties curves formed by the intersection of a surface with some plane containing a point on the surface. In particular, if one wants to study the curvature properties of an oriented surface $(\Sigma,\, \mathbf{N})$ at some point $\mathbf{p} \in \Sigma$, one might consider the curves formed by intersecting Σ with all planes containing the normal line to $\mathbf{p}$ and attempt to describe their curvatures. In fact, this approach leads directly to the basic notions of curvature for oriented surfaces in $\mathbf{R}^3$.

It will be convenient to review some concepts from the first unit of the course. If we are given a regular smooth curve $\mathbf{y}$ in $\mathbf{R}^3$ such that $\mathbf{y}(0) = \mathbf{p}$, then we shall let s denote the modified arc length parametrization such that $s(0) = 0$ and $s'(t) = |\mathbf{y}'(t)|$. Then one has the unit tangent vector function

$$\mathbf{T} \;=\; \frac{d\,\mathbf{y}}{d\,s} \;=\; \frac{\mathbf{y}'(t)}{s'(t)}$$

and the associated curvature vector function

$$\mathbf{k}(s) \;=\; \frac{d\,\mathbf{T}}{d\,s} \;=\; \frac{\mathbf{T}\,'(t)}{s'(t)}$$

that is perpendicular to $\mathbf{T}$ and whose magnitude is equal to the curvature of of $\mathbf{y}$ at a given parameter value.

Definition. Let $\mathbf{y}$ be a smooth curve in Σ such that $\mathbf{y}(0) = \mathbf{p}$; if $\mathbf{X}$ is a regular smooth parametrization of Σ at $\mathbf{p}$ then we may write $\mathbf{y}(t) = \mathbf{X}(\,u(t),\, v(t)\,)$ for suitable smooth functions u and v, at least if t is sufficiently close to 0. The *normal curvature vector* $\mathbf{k}_n$ for the curve is then given by

$$\mathbf{k}_n(s) \;=\; \langle \mathbf{k}(s),\, \mathbf{N}(\,\mathbf{y}(s)\,)\rangle \cdot \mathbf{N}(\,\mathbf{y}(s)\,)$$

and the *normal curvature* of $\mathbf{y}$ with respect to Σ is given by

$$\kappa_n \;=\; \langle \mathbf{k},\, \mathbf{N}\,\rangle\ .$$

The normal curvature vector and the normal curvature are related by the equation $\mathbf{k}_n = \kappa_n \cdot \mathbf{N}$.

IMPORTANT SPECIAL CASE. Suppose that we are given a curve $\mathbf{y}$ defined as the intersection of Σ with the plane through $\mathbf{p}$ that contains the normal line to Σ and the tangent line through $\mathbf{p}$ that is parallel to the nonzero vector $\mathbf{v} \in T_{\mathbf{p}}(\Sigma)$. Then the curvature vector for $\mathbf{y}$ at parameter value is perpendicular to $\mathbf{v}$ and lies in the plane containing this vector and $\mathbf{N}(\mathbf{p})$, and accordingly $\mathbf{k}(0)$ is a scalar multiple of $\mathbf{N}(\mathbf{p})$. In this case *the absolute value of the normal curvature is the ordinary curvature* of $\mathbf{y}$ at parameter value 0.

The following crucial result allows us to describe the normal curvature in relatively familiar terms.

MEUSNIER'S THEOREM. *The normal curvature vector* $\mathbf{k}_n$ *and the normal curvature* κ_n *at* $\mathbf{p}$ *only depend upon* $\mathbf{y}'(0) = \mathbf{w}$*, and in fact we have the formula*

$$\kappa_n \;=\; \frac{\mathbf{II}(\mathbf{w},\, \mathbf{w})}{\mathbf{I}(\mathbf{w},\, \mathbf{w})}\ .$$

Of course, this is just the sort of expression that we considered at the end of the previous section.

Historical footnote. Born on June 19, 1754, at Tours, France, JEAN BAPTISTE MARIE MEUSNIER is known for his ideas on designing airships and his career as a military officer as well as his results on the differential geometry of surfaces. He was a student of G. Monge at the École Royale du Génie in Mézières, and he was the first person to envision an elongated airship as an alternative to a spherical balloon. His suggestion of an elliptical-shaped airship was advanced in 1784, just weeks after the first flights of hot air balloons by the Montgolfier brothers. Henri Giffard adopted much of Meusnier's design in his first successful powered airship. Meusnier also played a key role in the organization of the army of the First French Republic; he was severely wounded during a battle between the French and Prussians at Cassel (near Mainz, Germany), and died on June 13, 1793.

Proof of Meusnier's Theorem. We know that the unit tangent vector $\mathbf{T}$ to the curve is perpendicular to the unit normal vector $\mathbf{N}$ to the surface because the curve lies in the surface. Differentiating both sides of the expression $0 = \mathbf{T} \cdot \mathbf{N}$ and applying the Leibniz Rule, we see that

$$0 \;=\; \frac{d\,\mathbf{T}}{d\,t} \cdot \mathbf{N} \;+\; \mathbf{N} \cdot \frac{d\,\mathbf{T}}{d\,t}$$

This leads to the following string of equations:

$$\kappa_n \;=\; \mathbf{k} \cdot \mathbf{N} \;=\; \frac{d\,\mathbf{T}}{ds} \cdot \mathbf{N} \;=\; \frac{1}{s'(t)} \cdot \left(\frac{d\,\mathbf{T}}{d\,t} \cdot \mathbf{N} \right) \;=\; -\,\frac{1}{s'(t)} \cdot \left(\mathbf{T} \cdot \frac{d\,\mathbf{N}}{d\,t} \right) \;=$$

$$-\,\frac{1}{s'(t)^2} \cdot \left(\frac{d\,\mathbf{y}}{d\,t} \cdot \frac{d\,\mathbf{N}}{d\,t} \right) \;=\; -\,\frac{1}{|\mathbf{y}'(t)|^2} \cdot \left(\frac{d\,\mathbf{y}}{d\,t} \cdot \frac{d\,\mathbf{N}}{d\,t} \right)$$

and by the Chain Rule the last expression is equal to

$$-\,\frac{(u'\,\mathbf{X}_u \;+\; v'\,\mathbf{X}_v) \cdot (u'\,\mathbf{N}_u \;+\; v'\,\mathbf{N}_v)}{|u'\,\mathbf{X}_u \;+\; v'\,\mathbf{X}_v|^2} \;.$$

The denominator of this expression is equal to the First Fundamental Form at $(\mathbf{w},\, \mathbf{w})$. Furthermore, the numerator is equal to the inner product of $\mathbf{w}$ and $[D\,\mathbf{N}(\mathbf{p})](\mathbf{w})$, or equivalently the negative of the value of the Second Fundamental Form at $(\mathbf{w},\, \mathbf{w})$. It follows that κ_n is the quotient of the Second Fundamental Form by the First evaluated at $(\mathbf{w},\, \mathbf{w})$.∎

By the results of the preceding section, the normal curvatures attain maximum and minimum values, these are realized at the eigenvectors of $-D\,\mathbf{N}(\mathbf{p})$, and the values are equal to the eigenvalues of $D\,\mathbf{N}(\mathbf{p})$. The average of these eigenvalues is called the *mean curvature* and the product is called the *Gaussian curvature*. Classically these quantities are denoted by H and K respectively.

LOCAL FORMULAS. *If the oriented surface* $(\Sigma,\, \mathbf{N})$ *is given by a regular smooth parametrization* $\mathbf{X}$*, then the mean and Gaussian curvatures* H *and* K *are given by the following formulas:*

$$H \;=\; \frac{e\,G \,-\, 2\,f\,F + g\,E}{2\,(E\,G - F^2)} \;.$$

$$K \;=\; \frac{e\,g - f^2}{E\,G - F^2}$$

Proof. These are immediate consequences of the trace and determinant formulas at the end of the previous section.∎

Once again we shall consider our standard examples and describe their mean and Gaussian curvatures. For the plane, we know that the Second Fundamental Form is identically zero, and therefore it follows that both the mean and Gaussian curvatures are zero everywhere. Suppose now that we consider the sphere defined by the equation $x^2 + y^2 + z^2 - r^2 = 0$ where $r > 0$. In this case the unit normal is given by $(x/r,\, y/r,\, z/r)$, so the Second Fundamental Form is just $-1/r$ times the First Fundamental Form. It follows that the normal curvature of every smooth curve through a point on the sphere is equal to $-1/r$, which in turn means that the mean curvature is equal to $-1/(2\,r)$ at each point and the Gaussian curvature is equal to $1/r^2$ at each point. Consider next the cylinder defined by $x^2 + y^2 = 1$. We noted that the eigenvalues of the map $D\,\mathbf{N}$ in this case were equal to 0 and 1, and therefore the mean and Gaussian curvatures in this case are equal to $\frac{1}{2}$ and 0 respectively. Finally, one can use our previous discussion for the hyperbolic paraboloid to find its mean and Gaussian curvatures at each point. One important new feature is that these quantities are no longer constants. We shall not go through all the details of this case but simply note that the mean and Gaussian curvatures for the hyperbolic paraboloid are negative at every point.

Note. The Gaussian curvature can in fact be defined for nonorientable surfaces. This is based upon the following observations:

(*i*) Locally the surface is given by a regular parametrization.

(*ii*) Surfaces given by a single regular parametrization are orientable, and locally they have exactly two orientations, one of which is the negative of the other.

(*iii*) One can use the preceding methods to compute the Gaussian curvature near a point $\mathbf{p}$, and the value is the same for both orientations near $\mathbf{p}$ essentially because the product of the eigenvalues for a diagonalizable 2×2 matrix A is the same as the product for $-A$.

In particular, this means that we may define the Gaussian curvature on the Möbius strip even though there is no globally defined smooth unit normal. These considerations also show that if Σ has an orientation, then the Gaussian curvature does not depend upon the specific choice of orientation.

Interpreting the sign of Gaussian curvature

As a first step to understanding the meaning of curvature for surfaces, it is important to consider the implications for the shape of the surface Σ if the Gaussian curvature is positive, negative or zero at a point. Since Gaussian curvature is continuous, if it is positive or negative at $\mathbf{p}$ then it is also positive or negative at all points close to $\mathbf{p}$, but if the Gaussian curvature is zero at $\mathbf{p}$ then one does not expect to draw any conclusion at all about the nonnegativity or nonpositivity of the Gaussian curvature near $\mathbf{p}$, and in fact we shall give examples to show that the shape of a surface near a point can vary significantly if the Gaussian curvature at $\mathbf{p}$ is zero.

The first step in using Gaussian curvature to obtain a rough idea of the shape near a point $\mathbf{p}$ is to consider the curves formed by intersection the surface with a plane $\mathbf{Q}$ containing the normal line M through $\mathbf{p}$. Let L be the tangent line to the intersection curve at $\mathbf{p}$, let $\mathbf{v}$ be a unit vector parallel to L, and let $\mathbf{Q}_+$ be the "positive" side of L in $\mathbf{Q}$ consisting of all points $\mathbf{x} \in \mathbf{Q}$ such that

$$(\mathbf{x} - \mathbf{p}) \cdot \mathbf{N}(\mathbf{p}) \quad > \quad 0 \; .$$

If the curve through $\mathbf{p}$ determined by $\Sigma \cap \mathbf{Q}$ has positive normal curvature at $\mathbf{p}$, this means that the center of the osculating circle at $\mathbf{p}$ is a point of M that lies on $\mathbf{Q}_+$, and in fact it follows that all points on the curve that are close to $\mathbf{p}$, except for $\mathbf{p}$ itself, also lie on $\mathbf{Q}_+$. Similarly, let $\mathbf{Q}_-$ denote the other side of L in $\mathbf{Q}$ consisting of all points for which $(\mathbf{x} - \mathbf{p}) \cdot \mathbf{N}(\mathbf{p})$ is negative, and suppose that the Gaussian curvature of the curve through $\mathbf{p}$ determined by $\Sigma \cap \mathbf{Q}$ is negative. Then it follows that the center of the osculating circle at $\mathbf{p}$ is a point of M that lies on $\mathbf{Q}_-$, and in fact it follows that all points on the curve that are close to $\mathbf{p}$, except for $\mathbf{p}$ itself, also lie on $\mathbf{Q}_-$.

Suppose now that the Gaussian curvature at $\mathbf{p}$ is positive. This means that the maximum and minimum values for the normal curvatures of the intersection curves are nonzero and have the same sign. In this case it follows that all points of the intersection curves that are sufficiently close to $\mathbf{p}$ lie on one of the closed sides of the tangent plane that are determined by one of the inequalities $(\mathbf{x} - \mathbf{p}) \cdot \mathbf{N}(\mathbf{p}) \geq 0$ or $(\mathbf{x} - \mathbf{p}) \cdot \mathbf{N}(\mathbf{p}) \leq 0$. Furthermore, with the exception of $\mathbf{p}$ itself, all of the nearby points on such curves lie on the open sides defined by replacing $\leq$ and $\geq$ with strict inequalities. This corresponds to the notion of strict local convexity that was discussed in the exercises.

Before proceeding, it will be useful to set some notation. Given a nonzero tangent vector $\mathbf{w}$ at $\mathbf{p}$, let $\mathbf{B}(\mathbf{w})$ be equal to $\mathbf{w} \times \mathbf{N}(\mathbf{p})$; since $\mathbf{w}$ and $\mathbf{N}(\mathbf{p})$ are nonzero vectors that are perpendicular to each other, it follows that their cross product is nonzero and perpendicular to both of these vectors. The plane containing $\mathbf{p}$ with normal direction corresponding to $\mathbf{B}(\mathbf{w})$ will be denoted by $\mathbf{Q}(\mathbf{w})$, and its two open sides $\mathbf{Q}_\pm(\mathbf{w})$ may then be defined as in the previous paragraph.

Suppose now that the Gaussian curvature at $\mathbf{p}$ is negative. In this case one can choose nonzero tangent vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ at $\mathbf{p}$ such that the normal curvatures of the corresponding plane intersections $\mathbf{Q}(\mathbf{v}_1) \cap \Sigma$ and $\mathbf{Q}(\mathbf{v}_2) \cap \Sigma$ are positive in the first case and negative in the second. If β_i is the curve near $\mathbf{p}$ determined by the intersection $\mathbf{Q}(\mathbf{v}_i) \cap \Sigma$, then the centers of the osculating circles for β_1 and β_2 lie on $\mathbf{Q}(\mathbf{v}_1)_+$ and $\mathbf{Q}(\mathbf{v}_2)_-$ respectively, and similar statements hold for all points of β_1 and β_2 that are close to $\mathbf{p}$ except for $\mathbf{p}$ itself. A good model for this is the saddle surface defined by the equation

$$z \;=\; y^2 \;-\; x^2$$

near the origin. The tangent plane of this surface at the origin is the xy-plane, and the standard upward normal for the surface at the origin is the vector $(0, 0, 1)$. Direct calculation shows that the First Fundamental Form is

$$(1 \;+\; 4\,x^2)\,dx\,dx \;+\; 16\,x^2\,y^2 \;+\; (1 \;+\; 4\,y^2)\,dy\,dy \;.$$

Furthermore, the intersections of this surface with the yz- and xz-planes are the parabolas $z = y^2$ and $z = -x^2$ respectively. Away from the origin, these intersection curves lie in the open half planes determined by the strict inequalities $z > 0$ and $z < 0$ respectively. Incidentally, one can show directly that the saddle surface has negative Gaussian curvature at the origin using the formulas from Section IV.2 and this section as follows: By the formulas at the end of Section IV.2 we know that the Second Fundamental form is given by

$$\frac{2\,(dy\,dy \;-\; dx\,dx)}{\sqrt{1 \;+\; 4\,x^2 \;+\; 4\,y^2}}$$

and therefore the formula for the Gaussian curvature in terms of the Second Fundamental Form shows that

$$K \;=\; \frac{-4}{(1 \;+\; 4\,x^2 \;+\; 4\,y^2)^2} \;.$$

Thus the Gaussian curvature is negative at all points of this surface.

Note that the mean curvature must be nonzero if the Gaussian curvature is positive because the latter implies that the maximum and minimum values of the sectional curvature are both positive or both negative.

If the Gaussian curvature at $\mathbf{p}$ is zero, then one cannot draw many conclusions about the shape of the surface near $\mathbf{p}$. This is best seen using examples. There are two basic cases depending on whether the Second Fundamental Form is zero or nonzero. We shall begin by considering the second possibility. In this case the map $D\,\mathbf{N}(\mathbf{p})$ is not invertible but also nonzero, and therefore it has two eigenvalues, one of which is zero and one of which is nonzero. The cylinder defined by $x^2+y^2 = 1$ is one example of this sort. A typical point of this surface is the unit vector $(1, 0, 0)$, and its tangent plane is defined by the equation $x = 1$. In this case all the curves formed by intersecting the surface with planes containing the normal line at $(1, 0, 0)$, which is the x-axis, lie on the sides of the tangent plane determined by the inequality $x \leq 1$, and in fact all points on these curves except $(1, 0, 0)$ itself lie in the set determined by the strict inequality $x > 1$. However, it is also possible to describe other examples where $D\,\mathbf{N}(\mathbf{p})$ has rank 1 but the surface has points on both sides of the tangent plane. The graph of the function $z = x^2 + y^3$ at the origin is a specific example (look at the intersection with the yz-plane).

If the Gaussian curvature is zero and $D\,\mathbf{N}(\mathbf{p}) = 0$ then the mean curvature is also zero and the local behavior of the surface near $\mathbf{p}$ also cannot be determined without additional information. A plane is the simplest example of this type. However, there are also examples for which the surface is strictly locally convex near the point $\mathbf{p}$ and examples where the surface has points on both open sides of the tangent plane near $\mathbf{p}$. An example where strict local convexity holds is given by the graph of $f(x, y) = x^4 + y^4$ at the origin, where the tangent plane to the surface is merely the xy-plane. One can use the methods employed for the saddle surface to show that the Second Fundamental Form is zero at the origin. If one intersects this surface with a plane containing the normal line at the origin, which is the z-axis, then the resulting curves all lie on the side of the tangent plane defined by the inequality $z \geq 0$, and except for the origin itself all points of the curve lie on the open side where the strict inequality $z > 0$ holds. On the other hand, consider the *Monkey Saddle Surface* defined by the equation

$$z \;=\; x^3 \;-\; 3\,x^2\,y$$

at the origin. Once again the tangent plane is the xy-plane and the Second Fundamental Form is zero. Using cylindrical coordinates and simple trigonometric identities, one can rewrite the equation of the surface as $z = r^3 \cos 3\theta$, and from this one sees that the intersection of the surface with a plane containing a normal line has a parametrization of the form

$$\left(\, t\,\cos 3\,\theta_0,\; t\,\sin 3\,\theta_0,\; t^3 \,\right)$$

for some fixed real number θ_0. These curves are line in the xy-plane if θ_0 is an integral multiple of $\pi/3$. On the other hand, for other choices of θ_0 the points on this curve corresponding to $t > 0$ and $t < 0$ lie on the two opposite open sides of the tangent plane defined by $z > 0$ and $z < 0$. In some cases the points corresponding to parameter values $t > 0$ lie on the side defined by $z > 0$, while in other cases these points lie on the side defined by $z < 0$. The following online sites contain excellent (and in the second case interactive) pictures of the Monkey Saddle:

`http://astronomy.swin.edu.au/~pbourke/surfaces/monkey/`

`http://www.ma.umist.ac.uk/kd/geomview/monkeysad.html`

`http://www.ag.jku.at/digpics_en.html`

IV.5 : Special classes of surfaces

(do Carmo, §3–5)

In these notes particular attention has been given to understanding the main concepts in the differential geometry of surfaces for the objects encountered in analytic geometry and calculus, including quadric surfaces, surfaces of revolution and certain examples of ruled surfaces. Needless to say, mathematicians and scientists in related fields have also found numerous other examples of surfaces that are curious, interesting or important for one reason or another. The purpose of this section is to discuss a few additional examples beyond the usual ones from analytic geometry and calculus and also to comment further on the geometric interpretation of mean and Gaussian curvature for some standard examples that have not yet been considered. These and other examples are particularly useful in illustrating the sorts of geometric insights one can obtain by means of methods from ordinary and multivariable calculus.

Here are some online references that have particularly good collections of surface graphics:

`http://www.uib.no/People/nfytn/mathgal.htm`

`http://www.uta.edu/optics/sudduth/4d/the_main_gallery.htm`

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

`http://www.math.arizona.edu/~models/Ruled_Surfaces/`

Of course, there are also many very good illustrations in do Carmo, but the impact advances in computer technology since the publication of this book is clear (and the previously cited book by Gray goes into the uses of such technology for differential geometry in great detail).

Ruled surfaces

We begin with a very simple observation.

GAUSSIAN CURVATURE OF A RULED SURFACE. *If the surface Σ is given by a ruled parametrization in the sense of Section* III.2*, then its Gaussian curvature is nonpositive.*

The cylinder and plane are examples of ruled surfaces for which the Gaussian curvature is identically zero, and both the hyperboloid of one sheet and the hyperbolic paraboloid (saddle surface) are examples of ruled surfaces for which the Gaussian curvature is always negative.

Proof. We can do this without computing the Gaussian curvature explicitly. Suppose that we have a ruled parametrization

$$\mathbf{X}(u,v) \;=\; \mathbf{a}(u) + v \cdot \mathbf{b}(u)$$

where $\mathbf{a}'(u)$ is never zero and the vectors $\mathbf{a}'(u)$ and $\mathbf{b}(u)$ are always linearly independent. Then the space of tangent vectors at $\mathbf{X}(u,v)$ is spanned by the linearly independent vectors $\mathbf{a}'(u)$ and $\mathbf{b}(u)$.

Consider the curve through $\mathbf{p} = \mathbf{X}(u,v)$ formed by intersecting the tangent plane to $\mathbf{p}$ at that point with the unique plane that contains $\mathbf{X}(u,v)$ and whose normal line is parallel to the vector $\mathbf{N}(\mathbf{p}) \times \mathbf{b}(u)$. This intersection is locally given by the line through $\mathbf{p}$ that is parallel to $\mathbf{b}(u)$. Of course the curvature of this curve is equal to zero and therefore we know that there is one tangent direction at $\mathbf{p}$ for which the sectional curvature is zero. If the Gaussian curvature were

either positive or negative, then the sectional curvature would be nonzero in every direction, and therefore the Gaussian curvature cannot be positive at **p**.∎

Derivation of formulas for the mean and Gaussian curvature of a ruled surface are left to the reader as an exercise.

Ruled surfaces for which the Gaussian curvature is identically zero are called *developable surfaces*, and they have many significant properties. Further information on this topic may be found on pages 194 and 210 of do Carmo.

Surfaces of revolution

We shall derive formulas for the Gaussian curvature of a surface of revolution obtained by rotating a curve in the xy-plane about the x- and y-axes. In the first case we need to assume that the x-coordinates for all points on the curve are positive, and in the second we need to make a similar assumption regarding the y-coordinates. Our ultimate goal is to describe a surface of revolution whose Gaussian curvature is equal to -1 at each point.

Before setting up the computations it is worthwhile to consider some examples in order to have a rough idea about what the general formulas for Gaussian curvature can be expected to yield. If we take $h(x) = \sqrt{1 - x^2}$ and rotate it around the x=axis we obtain a portion of the unit sphere centered at the origin. This surface has Gaussian curvature equal to $+1$ at each point, and the second derivative of h is negative for $-1 < x < 1$. On the other hand, if we take $h(x) = \sqrt{1 + x^2}$ and rotate it around the x=axis we obtain a portion of the hyperboloid of one sheet defined by the equation $y^2 + z^2 - x^2 = 1$, which has negative Gaussian curvature; in this case the second derivative of h is positive everywhere. This and further experimentation suggest that the signs of the second derivative and the Gaussian curvature should be the opposites of each other.

Suppose now that we are given a regular smooth curve $\mathbf{c}(t) = (\, p(t),\, q(t)\,)$ where $q(t) > 0$ for all t. Then the regular surface formed by rotating this curve about the x-axis may be given using the parametrization

$$\mathbf{X}(u, v) \;=\; \big(p(u),\, q(u)\,\cos v,\, p(u)\,\sin v\big)\ .$$

In order to compute the coefficients of the fundamental forms we need to find the partial derivatives $\mathbf{X}_1$ and $\mathbf{X}_2$ and the unit normal associated to the parametrization, which has the form

$$\mathbf{N} \;=\; \frac{1}{|\Omega|}\cdot\Omega$$

where $\Omega = \mathbf{X}_1 \times \mathbf{X}_2$. Here are the relevant formulas:

$$\begin{aligned}
\mathbf{X}_1 &= (\, p',\, q'\,\cos v,\, q'\,\sin v\,) \\
\mathbf{X}_2 &= (\, p,\, -q\,\sin v,\, q\,\cos v\,) \\
\Omega &= (\, q\,q',\, -p'\,q\,\cos v,\, -p'\,q\,\sin v\,) \\
|\Omega| &= q\,\sqrt{(p')^2 \,+\, (q')^2}
\end{aligned}$$

Given these formulas we see that the First Fundamental Form has coefficients $E = (p')^2 + (q')^2$, $F = 0$ and $G = q^2$.

Similarly, the second partial derivatives of $\mathbf{X}$ are given as follows:

$$\begin{aligned}
\mathbf{X}_{1,1} &= (p'',\, q''\cos v,\, q''\sin v) \\
\mathbf{X}_{1,2} &= \mathbf{X}_{2,1} = (0,\, -q'\sin v,\, -q'\cos v) \\
\mathbf{X}_{2,2} &= (0,\, -q\cos v,\, -q\sin v)
\end{aligned}$$

The corresponding inner products with Ω are given by

$$\begin{aligned}
\Omega \cdot \mathbf{X}_{1,1} &= p''\,q\,q' \;-\; p'\,q\,q'' \\
\Omega \cdot \mathbf{X}_{1,2} &= 0 \\
\Omega \cdot \mathbf{X}_{2,2} &= p'\,q^2
\end{aligned}$$

and therefore the coefficients of the Second Fundamental Form are given as follows:

$$\begin{aligned}
e &= \frac{p''\,q' \;-\; p'\,q''}{\sqrt{(p')^2 \;+\; (q')^2}} \\
f &= 0 \\
g &= \frac{p'\,q}{\sqrt{(p')^2 \;+\; (q')^2}}
\end{aligned}$$

These computations yield the following formula for the Gaussian curvature:

$$K = \frac{e\,g - f^2}{E\,G - F^2} = \frac{e\,g}{E\,G} =$$

$$\frac{(\,p''\,q \;-\; p'\,q'')\,p'\,q}{[(p')^2 \;+\; (q')^2]^2\,q^2} = \frac{(\,p''\,q \;-\; p'\,q'')\,p'}{[(p')^2 \;+\; (q')^2]^2\,q}$$

SPECIAL CASES. Suppose first that $p(t) = t$ so that the curve is simply the graph of a smooth function. Then the formula reduces to

$$K = \frac{-q''}{[1 \;+\; (q')^2]^2 \cdot q}$$

and therefore the signs of K and q'' are opposite, exactly as our examples suggested.

Suppose now that we assume that $|\mathbf{c}'(t)| \equiv 1$, so that $(p')^2 \;+\; (q')^2 = 1$. If we differentiate this with respect to u we obtain the equation $p'\,p'' \;+\; q'\,q'' \;=\; 0$, and thus we may use these equations to rewrite the Gaussian curvature as $-q''/q$. When $\mathbf{c}$ gives the standard parametrization of the unit circle with $p(t) = \cos t$ and $q(t) = \sin t$, this gives another proof that the Gaussian curvature of the unit sphere is equal to 1, at least at all points except perhaps $(\pm\,1, 0, 0)$.

Gaussian curvature of the torus. The preceding also allows us to compute the Gaussian curvature of the torus given given by revolving the circle with equation

$$x^2 \;+\; (y-2)^2 \;=\; 1$$

about the x-axis. In this case parametric equations for the circle are given by $p(t) = \cos t$ and $q(t) = 2 + \sin t$, and by the formula given above the Gaussian curvature is equal to

$$\frac{-q''(t)}{q(t)} = \frac{\sin t}{2 \;+\; \sin t}\,.$$

This quantity is positive if $t \in (0, \pi)$, zero if $t = 0,\ \pi,\ 2\pi$, and negative if $t \in (\pi,\ 2\,\pi)$. Visually, it is positive on the piece of the surface obtained by revolving the upper semicircle about the x-axis, negative on the piece obtained by revolving the lower semicircle about the x-axis, and zero on the circles obtained by revolving the points $(\pm, 1,\ 1)$ about the x-axis. Note that the Second Fundamental Form is nonzero at all points where the Gaussian Curvature is equal to zero.

The tractrix and pseudosphere

We shall now apply the preceding calculuations to find a surface of revolution whose Gaussian curvature is equal to a negative constant. The standard example of this sort is the **pseudosphere**, and it is obtained by revolving a curve known as the **tractrix** around the x-axis.

From a physical perspective the tractrix is given as follows: Suppose that a person is initially standing at the origin in $\mathbf{R}^2$ and is holding a tightly stretched leash with a dog on the other end at $(0, a)$ where $a > 0$. Now suppose that the person begins walking in the positive direction along the x-axis and the dog's path is such that the leash is tightly stretched at each point. If $\mathbf{D}(t)$ and $\mathbf{P}(t)$ denote the positions of the dog and person at time t, these conditions translate into the following mathematical conditions:

(1) The line of the leash is the tangent line of the path taken by the dog.

(2) The distance between the dog and person is always equal to a.

(3) If parametric equations for the dog's path are given by $(\,u(t),\, v(t)\,)$, then both $u(t)$ and $v(t)$ are positive while their derivatives satisfy $u'(t) \ > \ 0 \ > \ v'(t)$.

Here are some online graphics, including one that is animated:

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

`http://bradley.bradley.edu/~delgado/122/Tractrix.pdf`

`http://www.amherst.edu/~amcastro/MathMedia/galleries/Curves/Tractrix.html`

With the information given above we may derive parametric equations for the tractrix as follows: The position of the person $\mathbf{P}(t)$ on the x-axis is the intersection of that line with the tangent line, which may be parametrized as

$$\mathbf{L}(s) \ = \ = \ \ \mathbf{D}(t) + s\,\mathbf{D}'(t)$$

with the x-axis. If s is the parameter value at which this line meets the x-axis, then the mathematical conditions imply that $v(t) + s\,v'(t) = 0$, $|s\,\mathbf{c}'(t)| = a$ and $s\,u'(t) = \sqrt{a^2 - v(t)^2}$. Combining these equations, we conclude that

$$v(t)\,u'(t) \ \ = \ \ -s\,v'(t)\,u'(t) \ \ = \ \ -v'(t)\sqrt{a^2 - v(t)^2}\ .$$

Dividing these by the nonzero number $v(t)$ yields the differential equation

$$u' \ \ = \ \ \frac{\sqrt{a^2 - v^2}\cdot v'}{v}\ .$$

As usual, some initial conditions are needed in order to solve such equations uniquely. In our situation we know that

$$\lim_{t \to +\infty} \ u(t) \ \ = \ \ +\infty \quad \text{and} \quad \lim_{t \to +\infty} \ v(t) \ \ = \ \ +0\ .$$

If we make the change of variables $y = a \sin\theta$ then standard antidifferentiation formulas from integral calculus show that

$$x \;=\; a\left(\ln\tan(\theta/2) \;+\; \cos\theta\right) \;+\; C$$

and since the limit of y as $\theta \to \frac{1}{2}\pi$ is equal to a, it follows that the limit of x as $\theta \to \frac{1}{2}\pi$ is equal to 0. If we substitue this into thre right hand side of the formula for x, we see that the constant of integration C must be equal to zero. This means that one can describe the tractrix analytically by means of the parametrization

$$\left(\, a\left(\ln\tan(\theta/2) \;+\; \cos\theta\right),\, a\,\sin\theta\,\right) .$$

Note that as θ goes from 0 to $\pi/2$ this traces out the curve in the **reverse** direction from the one considered originally; the limiting values at $\theta = 0$ and $\pi/2$ may be viewed as $(+\infty,\, 0)$ and $(0,\, a)$ respectively.

To find the Gaussian curvature of the pseudosphere we may now substitute the coordinates for this parametrization into the general formulas given before:

$$p(\theta) \;=\; a\left(\ln\tan(\theta/2) \;+\; \cos\theta\right)$$

$$q(\theta) \;=\; a\,\sin\theta$$

If one then simplifies the resulting expression using standard differentiation rules and trigonometric identities, the conclusion is that $K = -1/a^2$ at all points of the pseudosphere. Thus the latter is indeed the desired *surface of revolution with prescribed constant negative Gaussian curvature.*

Constant Gaussian curvature

Classical Euclidean and Noneuclidean geometry have a natural interpretation in differential geometry as spaces of constant curvature. We have already seen that the plane and the sphere are spaces that have constant curvature, with the constant value equal to zero in the planar case and a positive number in the spherical case. In classical geometry it is either implicitly or explicitly assumed that the spaces have translational symmetry — given two points one can find a rigid motion sending one point to the other. From the perspective of differential geometry, this corresponds to an assumption that the Gaussian curvatures at every pair of points are the same, or equivalently that the Gaussian curvature is constant. The structure of surfaces with constant Gaussian curvature has been a central topic in differential geometry throughout its history, and there are important results which imply that all surfaces with constant curvature are very closely related to the fundamental examples; namely, the plane in the case of zero curvature, the sphere in the case of positive curvature, and the Noneuclidean plane discussed at the end of Section III.4. A fundamental theorem of D. Hilbert shows that one does not have a nice realization of the latter as a surface in $\mathbf{R}^3$, but the pseudosphere provides a good model in $\mathbf{R}^3$ for a small portion of this object (more precisely, if one removes the copy of the tractrix corresponding to $v = \pi$, then the remaining portion of the pseudosphere is metrically equivalent to a region in the Noneuclidean plane).

Minimal surfaces

A surface is said to be a *minimal surface* if its mean curvature is identically zero. This condition is simple and analogous to the conditions for constant Gaussian curvature, but none of this explains the reason for the name. From the viewpoint of local differential geometry, a minimal surface is is one that is equally bent in all directions so as to have zero average curvature just like a plane or the surface $z = x^2 - y^2$ at the origin, but in contrast to the latter one wants this property at every point of the surface. Aside from the plane, two basic examples of such surfaces are the catenoid and helicoid that are discussed on pages 202–205 of do Carmo.

Minimal surfaces are so named because of their connection to the following natural question: *Given a closed curve* Γ*, find the surface of least area that is bounded by* Γ. This question is known as *Plateau's Problem* and it is named after the physicist who noted that such surfaces may be realized physically by soap films that are bounded by the given curve.

As noted on page 197–199 of do Carmo, surfaces of least area must have mean curvatures that are identically zero, and the discussion on those pages provides strong evidence for this, at least in some relatively elementary situations. However, a precise mathematical formulation of the least area problem in a reasonably general context turns out to be extremely nontrivial.

Minimal surfaces have important relations to the theory of functions of a complex variable and partial differential equations; the most basic aspects of this are described on pages 201–202 of do Carmo. The study of minimal surfaces has had a strong impact on both geometry and analysis, in may cases leading to results on questions that at first doe not seem to have any relation to the least area problem.

Numerous examples of minimal surfaces have been discovered or constructed over the past two hundred years, and advances in computer technology during the past two dozen years have led to striking new insights, yielding unexpected new types of such surfaces whose existence was first suggested by computer graphics and later confirmed by rigorous mathematical proofs (but not all potential examples arising from computer graphics turned out to be minimal surfaces!). Here are some online references that discuss minimal surfaces, with many illustrations and more information on advances that have taken place during the past two dozen years:

`http://ctouron.freeshell.net/personal/costa/background.html`

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

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

`http://www.indiana.edu/ minimal/toc.html`

`http://www.zib.de/polthier/booklet/intro.html`

`http://www.csuohio.edu/math/oprea/soap/soap.html`

A book by J. Oprea, *Differential Geometry and Its Applications*, Second Edition (Prentice-Hall, 2003, ISBN: 0-13-065246-6), contains a detailed and current account of minimal surfaces at the undergraduate textbook level.

IV.6 : Compatibility equations, *Theorema Egregium*

(do Carmo, §4–3)

One of the most far-reaching results on the differential geometry of surfaces is that the Gaussian curvature of a surface can be expressed entirely in terms of the First Fundamental Form:

GAUSS' THEOREMA EGREGIUM. *If* $\mathbf{X}$ *is a* $1-1$ *regular parametrization such that the First Fundamental Form is given by*

$$E(u,v)\,du\,du \;+\; 2\,F(u,v)\,du\,dv \;+\; G(u,v)\,dv\,dv$$

and $K(u,v)$ *is the Gaussian curvature function, then the Gaussian curvature depends only upon the coeffieicnts of the First Fundamental Form of the surface and their partial derivatives.*

In contrast, the plane and cylinder have the same First Fundamental Form but different mean curvatures.

At the end of Section III.4 we discussed generalizations of the First Fundamental Form known as Riemannian metrics. One can use the formula above to define Gaussian curvature with respect to an arbitrary Riemannian metric regardless of whether it comes from a First Fundamental Form. This is an important step in formulating general notion of curvature in differential geometry that can be used in many different contexts and for an enormous number of purposes.

Discussion of the proof

The results of this section and the next are based upon an analysis of the second partial derivatives of a regular parametrization $\mathbf{X}$. In some sense this is analogous to the idea behind the Frenet-Serret Formulas for curves; one writes out the various derivatives as linear combinations of simpler objects and looks for useful interrelationships. In the case of curves, the Frenet Trihedron provided a useful basis for $\mathbf{R}^3$ at each point of the curve. For surfaces given by regular parametrizations, the corresponding useful basis is given by the partial derivatives $\mathbf{X}_1$ and $\mathbf{X}_2$ together with the unit normal vector $\mathbf{N}$, which may be viewed as $\mathbf{X}_1 \times \mathbf{X}_2$ normalized to have unit length. One major difference with the theory for curves is that these bases are usually not orthonormal, but this turns out to be a relatively minor issue that can be addressed directly using linear algebra as in the final portion of Section IV.3.

If one writes out the partial derivatives of $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{N}$ with respect to the u (first) and v (second) variables and uses the earlier computations involving the First and Second Fundamental Forms, one obtains the following sorts of formulas in which the quantities $\Gamma^i_{j,k}$ are smooth functions of u and v and are called *Christoffel symbols of the second kind*; the terminology is chosen to be consistent with concepts in tensor analysis (see the bottom of page 213 in the Schaum's Outline Series book on differential geometry for further information).

$$
\begin{aligned}
\mathbf{X}_{1,1} &= \Gamma^1_{1,1}\,\mathbf{X}_1 \,+\, \Gamma^2_{1,1}\,\mathbf{X}_2 \,+\, e\,\mathbf{N} \\
\mathbf{X}_{1,2} &= \Gamma^1_{1,2}\,\mathbf{X}_1 \,+\, \Gamma^2_{1,2}\,\mathbf{X}_2 \,+\, f\,\mathbf{N} \\
\mathbf{X}_{2,2} &= \Gamma^1_{2,2}\,\mathbf{X}_1 \,+\, \Gamma^2_{2,2}\,\mathbf{X}_2 \,+\, g\,\mathbf{N} \\
\mathbf{N}_1 &= \beta^1_1\,\mathbf{X}_1 \,+\, \beta^2_1\,\mathbf{X}_2 \\
\mathbf{N}_2 &= \beta^1_2\,\mathbf{X}_1 \,+\, \beta^2_2\,\mathbf{X}_2
\end{aligned}
$$

It is convenient to define $\Gamma^i_{2,1} = \Gamma^i_{1,2}$ for $i = 1,\, 2$ so that $\Gamma^i_{j,k}$ is defined for $1 \le i,\, j,\, k \le 2$ and satisfies $\Gamma^i_{k,j} = \Gamma^i_{j,k}$.

Using the methods described in the last part of Section IV.3 one can solve for β^i_j in terms of the coefficients of the First and Second Fundamental Forms:

$$
\begin{aligned}
\beta_{1,1} &= \frac{f\,F \,-\, e\,G}{E\,G \,-\, F^2} \\
\beta_{2,1} &= \frac{e\,F \,-\, f\,E}{E\,G \,-\, F^2} \\
\beta_{1,2} &= \frac{f\,F \,-\, f\,G}{E\,G \,-\, F^2} \\
\beta_{1,2} &= \frac{f\,F \,-\, g\,E}{E\,G \,-\, F^2}
\end{aligned}
$$

If one substitutes these into the equations for $\mathbf{N}_1$ and $\mathbf{N}_2$ one obtains the *Weingarten equations*. Computing the Christoffel symbols is more difficult. The following formulas are derived in Problem 10.3 on page 216 of the Schaum's Outline Series review of differential geometry that was cited previously:

$$
\begin{aligned}
\Gamma^1_{1,1} &= \frac{G\,E_1 \,-\, 2\,F\,F_1 \,+\, F\,E_2}{2\,(E\,G \,-\, F^2)} \\
\Gamma^1_{1,2} &= \frac{G\,E_2 \,-\, F\,G_1}{2\,(E\,G \,-\, F^2)} \\
\Gamma^1_{2,2} &= \frac{2\,G\,F_2 \,-\, G\,G_1 \,+\, F\,G_2}{2\,(E\,G \,-\, F^2)} \\
\Gamma^2_{1,1} &= \frac{2\,E\,F_1 \,-\, E\,E_2 \,+\, F\,E_1}{2\,(E\,G \,-\, F^2)} \\
\Gamma^2_{1,2} &= \frac{E\,G_1 \,-\, F\,E_2}{2\,(E\,G \,-\, F^2)} \\
\Gamma^2_{2,2} &= \frac{E\,G_1 \,-\, 2\,F\,F_2 \,+\, F\,G_1}{2\,(E\,G \,-\, F^2)}
\end{aligned}
$$

It is important to note that the Christoffel symbols depend only upon the coefficients of the First Fundamental Form and their first partial derivatives.

The most direct approach to proving Gauss' theorem is to continue by proving that

$$
K\,(E\,G \,-\, F^2)^2 \;=\; [\mathbf{X}_{1,1},\, \mathbf{X}_1,\, \mathbf{X}_2\,] \cdot [\mathbf{X}_{2,2},\, \mathbf{X}_1,\, \mathbf{X}_2\,] \,-\, [\mathbf{X}_{1,2},\, \mathbf{X}_1,\, \mathbf{X}_2\,]^2 \;.
$$

This compuation is carried out in Problem 10.4 on page 217 of the Schaum's Outline Series on differential geometry. Further computations using the same methods then yield the identity

$$K\,(E\,G\;-\;F^2)^2 \;\;=\;\; \left(\,F_{1,2}\;-\;\tfrac{1}{2}\,E_{2,2}\;-\;G_{1,1}\,\right)\,+$$

$$\left|\begin{array}{ccc} 0 & F_2-\frac{1}{2}\,G_1 & \frac{1}{2}\,G_2 \\ \frac{1}{2}\,E_1 & E & F \\ F_1-\frac{1}{2}\,E_2 & F & G \end{array}\right| \;-\; \left|\begin{array}{ccc} 0 & \frac{1}{2}\,E_2 & \frac{1}{2}\,G_1 \\ \frac{1}{2}\,E_2 & E & F \\ \frac{1}{2}\,G_1 & F & G \end{array}\right|$$

which implies that K depends only upon the coefficients of the First Fundamental Form and their partial derivatives. It is an elementary exercise in partial differentiation to show that this equation is equivalent to the one in the statement of Gauss' theorem that is given above.∎

Another approach to deriving Gauss' theorem is given on pages 231–235 in Section 4–3 of do Carmo. This alternate approach also has other implications, and it will be discussed in the final section of these notes.

Curvature and the First Fundamental Form

We have discussed the geometric significance of Gaussian curvature for a surface in $\mathbf{R}^3$ in terms of its First and Second Fundamental Forms. The *Theorema Egregium* provides a way of defining the Gaussian curvature entirely in terms of the First Fundamental Form, and consequently for riemannian metrics that are not necessarily realizable by surfaces in $\mathbf{R}^3$. One is therefore led to natural questions about interpreting the Gaussian curvature entirely in terms of metrical properties directly given the First Fundamental Form without using auxiliary objects such as normal lines or osculating circles. We shall describe one interpretation of positive and negative Gaussian curvature at a point entirely in metric terms; if the Gaussian curvature is equal to zero the situation is more complicated, but if the Gaussian curvature is identically zero then we shall give a similar interpretation.

Given a Riemannian metric

$$E(u,v)\,du\,du \;+\; 2\,F(u,v)\,du\,dv \;+\; G(u,v)\,dv\,dv$$

and a parametrized regular, piecewise smooth curve in a connected domain $U \subset \mathbf{R}^2$ on which the metric is defined, one can define the **length** of the curve by the formula

$$\int_a^b \;\sqrt{\,E(u,v)\,u'(t)^2 \;+\; 2\,F(u,v)\,u'(t)\,v'(t)v \;+\; G(u,v)\,v'(t)^2}\;\,dt$$

where the curve is defined on the interval $[a,b]$. The positivity condition on the coefficients E, F and G for a Riemannian metric imply that the expression inside the square root sign is always positive for regular smooth curves. One would like to define the *distance between two points* with respect to this metric as the greatest lower bound of the lengths of all regular piecewise smooth curves joining the points.

Two questions immediately arise. First of all, one needs to show that the lengths of curves joining two distinct points are bounded from below by a positive constant; in other words, if $\mathbf{p}$ and $\mathbf{q}$ are distinct points of a surface then it is not possible to find a sequence of piecewise smooth regular curves $\mathbf{y}_n$ joining them such that the length of $\mathbf{y}_n$ is less than $1/n$. Second, one would like to know if there is some curve for which the greatest lower bound is actually realized. Such a curve is called a *minimal geodesic*.

It is fairly easy to construct a somewhat artificial example where there is no curve of minimum length joining two points. Specifically, consider the surface given by removing the origin from the xy-plane. Then the greatest lower bound of the lengths of all piecewise smooth curves joining $(1,0,0)$ and $(-1,0,0)$ is equal to 2, which is the ordinary Euclidean distance, but there is no curve of length 2 joining these points that misses the origin. To see this, let $\mathbf{y}$ be a regular piecewise smooth curve joining the two points in question that is defined on $[a,b]$. Then there is some point $\xi \in (a,b)$ such that the first coordinate of $\mathbf{y}(\xi)$ is equal to some nonzero value, say c. It then follows by the Intermediate Value Theorem that the arc length of $\mathbf{y}$ must be greater than or equal to the length of the broken line curve joining $(1,0,0)$ to $(0,c,0)$ linearly and $(0,c,0)$ to $(-1,0,0)$ linearly. The length of this broken line curve is $2\sqrt{1+c^2}$, which is strictly greater than 1. Therefore there is no curve of shortest length joining the two points that lies completely inside the surface. One obvious feature of this example is that one can extend the given surface to a larger one (namely, the whole plane) in which there is a curve of minimum length joining the two points in question. In fact, one can construct examples for which one cannot add extra points to ensure that minimizing geodesics always exist, but such a construction would require a great deal of additional work. A natural candidate for a bad example is the graph of the function

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

which is defined for $(x,y) \neq (0,0)$ and cannot be extended to a function that is continuous at $(0,0)$.

In contrast to the preceding paragraph, it turns out that one can always find curves of minimum length joining a given point $\mathbf{p}$ to another point $\mathbf{q}$ provided $\mathbf{q}$ is sufficiently close to $\mathbf{p}$, and this fact has important implications to showing the the lengths of curves joining two distinct points are bounded from below by a positive constant.

EXISTENCE OF SHORT GEODESICS. *Suppose we are given a riemannian metric* $\mathbf{M}$ *on a connected domain in* $\mathbf{R}^2$*, and let* $\mathbf{p} \in U$*. Then there is an* $r > 0$ *such that* $|\mathbf{q} - \mathbf{p}| < r$ *implies that* $\mathbf{p}$ *and* $\mathbf{q}$ *can be joined by a regular piecewise smooth curve of least length, and this curve is in fact a regular smooth curve that lies entirely in the open disk with center* $\mathbf{p}$ *and radius* r*.*

Furthermore, given any nonzero vector $\mathbf{v} \in \mathbf{R}^2$ *there is a unique regular smooth curve curve* Γ *defined on an open interval* $(-h,h)$ *containing* 0 *such that* $\Gamma(0) = \mathbf{p}$*,* $\Gamma'(0) = \mathbf{v}$ *and* Γ *defines a curve of minimum length joining* $\mathbf{p}$ *to* $\Gamma(t)$ *for all* t *in the given interval* $(-h,h)$*.*

Finally, if $\delta \in (0,r)$ *and* $L(\mathbf{q})$ *denotes the length of the shortest curve joining* $\mathbf{p}$ *to* $\mathbf{q}$*, then the minimum value* $m(\delta)$ *of* $L(\mathbf{q})$ *over the circle defined by* $|\mathbf{q}| = \delta$ *is positive.*∎

The curves of least length in this result are called **minimizing geodesics**. It turns out that such curves are defined by second order differential equations, and this is the reason for the conclusion in the second paragraph.

COROLLARY. *If* Σ *is a surface and* $\mathbf{p}$ *and* $\mathbf{q}$ *are two points on* σ *that can be joined by a regular piecewise smooth curve on* Σ*, then the set of lengths for all such curves is bounded from below by a positive constant.*

Proof. To simplify the discussion we shall choose parametrizations for our regular piecewise smooth curves over some interval of the form $[0,a]$ such that $\Gamma(0) = \mathbf{p}$ and $\Gamma(a) = \mathbf{q}$. We need to find a positive lower bound for the length that does not depend upon the particular curve Γ.

Let $\mathbf{X}$ be a regular smooth parametrization for Σ at $\mathbf{p}$ that is 1–1, let $\mathbf{X}(\mathbf{p}_0) = \mathbf{p}$, and let $r > 0$ be as in the existence theorem stated above. There are two cases, depending upon whether the point $\mathbf{q} \in \Sigma$ has the form $\mathbf{X}(\mathbf{q}_0)$ for some $\mathbf{q}_0$ satisfying $|\mathbf{q}_0 - \mathbf{p}_0| < r$.

FIRST CASE. Suppose that $\mathbf{q}$ satisfies the condition in the preceding sentence, and let $s = |\mathbf{q}_0 - \mathbf{p}_0|$. If $\mathbf{y}$ is an arbitrary point of Σ having the form $\mathbf{X}(\mathbf{y}_0)$ for some $\mathbf{y}_0$ satisfying $|\mathbf{y}_0 - \mathbf{p}_0| < r$, then we shall define $g_0(\mathbf{y})$ to be equal to $|\mathbf{y}_0 - \mathbf{p}_0|$; the right hand side is well defined because the parametrization $\mathbf{X}$ is 1–1. This turns out to be a continuous function of $\mathbf{y}$. Likewise, if we define a real valued function g by setting $g(t) = \min\{s,\ g_0(\,\Gamma(t)\,)\,\}$ if $\Gamma(t)$ has the given special form, and $g(t) = s$ if $\Gamma(t)$ does not have this form, then g is continuous on the interval $[a, b]$ over which Γ is defined.

Since $g(a) = 0$ and $g(b) = s$, there must be a first parameter value t_0 such that $g(t_0) = s$. We claim that the image of the restricted curve $\Gamma|[0, t_0)$ lies in the image W of the disk of radius s centered at $\mathbf{p}_0$ under $\mathbf{X}$ (in fact a stronger statement is true but we shall not need this). This is true because if $\Gamma(t)$ does not lie in the image then $g(t) \geq s$ and we know that $g(t) < s$ if $t \in [0, t_0)$.

By the Intermediate Value Theorem there is a $t_1 \in (0, t_0)$ such that $g(t_1) = \frac{1}{2}s$; we know that the image of Γ restricted to $[0, t_1]$ lies in the set W described above, and therefore this restriction may be written as a composite $\mathbf{X} \circ \Gamma_1$ for some regular piecewise smooth curve Γ_1 which takes values in the disk of radius r centered at $\mathbf{p}_0$. We then have

$$\text{Length}\,(\Gamma|[0, t_1]) \quad = \quad \text{Length}_{\mathbf{M}}(\Gamma_1) \geq m(\tfrac{1}{2}\,s) \quad > \quad 0$$

on one hand and

$$\text{Length}\,(\Gamma|[0, t_1]) \quad \leq \quad \text{Length}(\Gamma)$$

on the other, which implies that the right had side is greater than or equal to the positive quantity $m(\frac{1}{2}s)$. This gives us our desired positive lower bound on the length of Γ which is independent of the curve Γ itself.

SECOND CASE. The argument is similar but not quite identical. We may define the function g exactly in the first case for an arbitrary s such that $0 < s < r$. In this case we know that $g(t) = s$ for some parameter value t because there is some value t such that $\Gamma(t)$ does **NOT** have the form $\mathbf{X}(\mathbf{y}_0)$ for some $\mathbf{y}_0$ satisfying $|\mathbf{y}_0 - \mathbf{p}_0| < r$. We can now proceed as before to find the least parameter value t_0 such that $g(t_0) = s$, and from this point on the argument is identical to the proof in the first case.∎

FUNDAMENTAL PROPERTIES OF DISTANCE FUNCTIONS. *Suppose that we have either a Riemannian metric* $\mathbf{M}$ *defined on a connected domain* U *in* $\mathbf{R}^n$ *or a geometric surface* Σ *in* $\mathbf{R}^3$ *such that each pair of points in* Σ *can be joined by a regular piecewise smooth curve in* Σ*, and let* $d_{\mathbf{M}}(\mathbf{x},\,\mathbf{y})$ *or* $d_{\Sigma}(\mathbf{x},\,\mathbf{y})$ *denote the greatest lower bound of the lengths of piecewise smooth curves joining* $\mathbf{x}$ *and* $\mathbf{y}$ *in* U *or* Σ*. Then this distance function* d *has the following basic properties:*

[1] *The distance* $d(\mathbf{x},\,\mathbf{y})$ *is nonnegative, and it is equal to zero if and only if* $\mathbf{x} = \mathbf{y}$.

[2] *For all* $\mathbf{x}$ *and* $\mathbf{y}$ *we have* $d(\mathbf{x},\,\mathbf{y}) = d(\mathbf{y},\,\mathbf{x})$*. and it is equal to zero if and only if* $\mathbf{x} = \mathbf{y}$.

[3] (TRIANGLE INEQUALITY) *For all* $\mathbf{x}$, $\mathbf{y}$ *and* $\mathbf{z}$ *we have*

$$d(\mathbf{x},\,\mathbf{z}) \quad \leq \quad d(\mathbf{x},\,\mathbf{y}) \ + \ d(\mathbf{y},\,\mathbf{z})$$

Sketch of proofs. The first statement follows from the immediately preceding discussion. To prove the second, not that if Γ is a regular piecewise smooth curve defined on $[a, b]$ joining $\mathbf{x}$ to $\mathbf{y}$ then $\Gamma^*(t) = \Gamma(b - t)$ defines a similar curve on $[a - b, 0]$ joining $\mathbf{y}$ to $\mathbf{x}$. This implies that $d(\mathbf{y},\,\mathbf{x}) \leq d(\mathbf{x},\,\mathbf{y})$. Reversing the roles of $\mathbf{x}$ and $\mathbf{y}$ yields the reverse inequality $d(\mathbf{y},\,\mathbf{x}) \geq d(\mathbf{x},\,\mathbf{y})$, and therefore the two quantities must be equal. Finally, to prove the third statement, let $\varepsilon > 0$ and choose suitable curves Γ_1 and Γ_2 such that Γ_i is defined on $[0, a_i]$, with Γ_1 joining $\mathbf{x}$ to $\mathbf{y}$ and Γ_2 joining $\mathbf{y}$ to $\mathbf{z}$, and the lengths of these curves satisfying

$$\text{Length}(\Gamma_1) \quad \leq \quad d(\mathbf{x},\,\mathbf{y}) \ + \ \frac{\varepsilon}{2}$$

$$\text{Length}(\Gamma_2) \quad \leq \quad d(\mathbf{y},\, \mathbf{z}) \;+\; \frac{\varepsilon}{2}\;.$$

Consider the curve formed by concatenating Γ_1 and Γ_2; specifically, let Γ be the curve defined on the interval $[0, a_1 + a_2]$ such that $\Gamma(t) = \Gamma_1(t)$ for $t \in [0, a_1]$ and $\Gamma(t) = \Gamma(t - a_1)$ for $t \in [a_1, a_1 + a_2]$. These piece together to form a regular piecewise smooth curve because the two formulas yield the same point at parameter value a_1. The length of this curve then given by

$$\text{Length}(\Gamma_1) \;+\; \text{Length}(\Gamma_2)$$

and hence we have the inequality

$$d((\mathbf{x},\, \mathbf{z}) \quad = \quad \text{Length}(\Gamma) \quad = \quad \text{Length}(\Gamma_1) \;+\; \text{Length}(\Gamma_2) \quad \leq \quad d(\mathbf{x},\, \mathbf{y}) \;+\; d(\mathbf{y},\, \mathbf{z}) \;+\; \varepsilon$$

for every $\varepsilon > 0$. In particular this implies that the expression on the left hand side cannot be greater than $d(\mathbf{x},\, \mathbf{y}) \;+\; d(\mathbf{y},\, \mathbf{z})$, and this is precisely the assertion in [3].∎

A METRIC INTERPRETATION OF CURVATURE. Suppose that we are given three points $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ in $\mathbf{R}^3$ that form the vertices of an isosceles triangle with vertex at $\mathbf{a}$; *i.e.*, we have $|\mathbf{b} - \mathbf{a}| = |\mathbf{c} - \mathbf{a}| = \ell > 0$. If θ is the angle between $\mathbf{b} - \mathbf{a}$ and $\mathbf{c} - \mathbf{a}$ then it is an elementary exercise in trigonometry to prove that $|\mathbf{c} - \mathbf{b}| = 2 \sin \frac{1}{2}\theta$. Roughly speaking, Gaussian curvature measures the extent to which this fails for riemannian metrics. The proof of this fact requires a considerable amount of machinery from Riemannian geometry, so we shall simply state the results here.

Since we are only concerned with metric behavior near a point, it will suffice to look at Riemannian metrics defined on an open disk centered at some point $\mathbf{p}$ in a connected domain $U \subset \mathbf{R}^2$. Let $\mathbf{M}$ be a Riemannian metric, and let $r > 0$ be so small that every point in the open disk of radius r centered at $\mathbf{p}$ can be joined to the later by a smooth curve of minimum length lying entirely inside this disk. Given two linearly independent vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbf{R}^2$, let $\theta_{\mathbf{M}}(\mathbf{p})$ be the angle between them computed with respect to the Riemannian metric:

$$\cos\left(\theta_{\mathbf{M}}(\mathbf{p})\right) \quad = \frac{\mathbf{M}_{\mathbf{p}}(\mathbf{v},\, \mathbf{w})}{\left(\mathbf{M}_{\mathbf{p}}(\mathbf{v},\, \mathbf{v})\right)^{1/2}\; \left(\mathbf{M}_{\mathbf{p}}(\mathbf{w},\, \mathbf{w})\right)^{1/2}}$$

Consider now the smooth geodesics which pass through $\mathbf{p}$ and have tangent vectors $\mathbf{v}$ and $\mathbf{w}$ at $\mathbf{p}$. We can find points on these geodesics that are some positive distances away from $\mathbf{p}$; if δ_0 is the minimum of the two distances, then for every $\ell \in (0, \delta_0]$ we can find points $\mathbf{x}$ and $\mathbf{y}$ on the respective geodesics such that the distances from $\mathbf{x}$ and $\mathbf{y}$ to $\mathbf{p}$ are both equal to ℓ (suppose that we have geodesics with the given tangent vectors defined on intervals $[0, a]$ and $[0, b]$ respectively; then by the Intermediate Value Theorem one can find points s_0 and t_0 in these intervals so that the lengths of the restrictions of the geodesics up to parameter values s_0 and t_0 are equal to ℓ). We then have the following relationships between the Gaussian curvature at $\mathbf{p}$ and the distance between $\mathbf{x}$ and $\mathbf{y}$ with respect to $\mathbf{M}$.

DISTANCE COMPARISON. *Suppose we are given everything as in the preceding discussion, and let K be the Gaussian curvature at $\mathbf{p}$.*

(i) If $K > 0$ then there is a $\delta_1 > 0$ such that if $\ell < \delta_1$ we have

$$d_{\mathbf{M}}(\mathbf{x},\, \mathbf{y}) \quad < \quad 2\,\ell\, \sin \theta_{\mathbf{M}}(\mathbf{p})\;.$$

(ii) If $K < 0$ then there is a $\delta_1 > 0$ such that if $\ell < \delta_1$ we have

$$d_{\mathbf{M}}(\mathbf{x},\, \mathbf{y}) \quad > \quad 2\,\ell\, \sin \theta_{\mathbf{M}}(\mathbf{p})\;.$$

(iii) If the Gaussian curvature is identically zero, then there is a $\delta_1 > 0$ such that if $\ell < \delta_1$ we have

$$d_{\mathbf{M}}(\mathbf{x},\, \mathbf{y}) \;\; = \;\; 2\,\ell\, \sin \theta_{\mathbf{M}}(\mathbf{p}) \;\blacksquare$$

These may be viewed as generalizations of standard trigonometric formulas from spherical geometry, Noneuclidean geometry in the sense of Bólyai and Lobachevsky, and classical Euclidean geometry respectively. Note that if we only know the Gaussian curvature is zero at $\mathbf{p}$ but we know nothing else about its behavior near $\mathbf{p}$, then these comparison results yield no information.

IV.7 : 7. Fundamental Theorem of Local Surface Theory

(do Carmo, §4–3, 4–Appendix)

The Frenet-Serret Formulas imply that curvature and torsion completely determine a curve locally provided on gives the initial position and unit tangent vector for the curve. There is a corresponding theorem for surfaces involving the coefficients E, F, G and e, f, g of the First and Second Fundamental Forms. However, these coefficient functions must satisfy some nontrivial restrictions. We have already noted that the matrix for the First Fundamental Form

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

must have positive eigenvalues, or equivalently that E and G as well as the determinant $E\,G - F^2$ must be positive. However, there are also other conditions that arise naturally from our basic assumptions that a local parametrizatin $\mathbf{X}$ have "sufficiently many" continuous partial derivatives. In particular, if we want $\mathbf{X}$ to have continuous third partial derivatives then we have equations of the form $\mathbf{X}_{1,1,2} = \mathbf{X}_{1,2,1} = \mathbf{X}_{2,1,1}$ and then we have equations of the form $\mathbf{X}_{2,2,1} = \mathbf{X}_{2,1,2} = \mathbf{X}_{1,2,2}$. If we combine these equations with the expansions of the second partial derivatives $\mathbf{X}_{i,j}$ in terms of Christoffel symbols and the Second Fundamental Form coefficients, we obtain the following three equations:

$$\begin{aligned}
e_2 - f_1 &= e\Gamma^1_{1,2} + f\left(\Gamma^2_{1,2} - \Gamma^1_{1,1}\right) - g\Gamma^2_{1,1} \\
f_2 - g_1 &= e\Gamma^1_{2,2} + f\left(\Gamma^2_{2,2} - \Gamma^1_{1,2}\right) - g\Gamma^2_{1,2} \\
e\,g - f^2 &= F\cdot\left[(\Gamma^2_{2,2})_1 - (\Gamma^2_{1,2})_2 + \Gamma^1_{2,2}\Gamma^2_{1,1} - \Gamma^1_{1,2}\Gamma^2_{1,1}\right] + \\
& E\cdot\left[(\Gamma^1_{2,2})_1 - (\Gamma^1_{1,2})_2 + \Gamma^1_{2,2}\Gamma^1_{1,1} + \Gamma^2_{2,2}\Gamma^1_{1,2} - \Gamma^1_{1,2}\Gamma^1_{1,2} - \Gamma^2_{1,2}\Gamma^1_{2,2}\right]
\end{aligned}$$

The first two of these are known as the *Codazzi-Mainardi Equations*. We note in passing that the third equation provides another demonstration of Gauss' Theorema Egregium; in fact, one important advantage of this proof is that it reflects the standard approach to curvature in the study of differential geometry for objects whose dimensions are greater than two.

The verifications of these formulas are carried out on pages 235–236 of do Carmo and in Problem 10.28 on page 224 of the Schaum's Outline Series book on differential geometry.■

FUNDAMENTAL THEOREM OF LOCAL SURFACE THEORY. *Let U be a connected domain in $\mathbf{R}^2$, and let E, F, G and e, f, g be smooth functions with sufficiently many continuous partial derivatives on U such that E, F and G satisfy the positive definiteness conditions given above and e, f and g satisfy the three compatibility conditions displayed above. Then for each $\mathbf{p}_0 \in U$, $\mathbf{p} \in \mathbf{R}^3$ and plane Π containing $\mathbf{p}$, there is a regular surface parametrization $\mathbf{X}$ defined on some open disk N about $\mathbf{p}_0$ such that the First and Second Fundamental Forms of $\mathbf{X}$ have coefficients equal to E, F, G and e, f, g respectively. This parametrization is locally unique up to a rigid motion of $\mathbf{R}^3$.*

The uniqueness proof is essentially a relatively lengthy argument involving the uniqueness of solutions of certain ordinary differential equations (see pages 236 and 311–314 of do Carmo or the argument following the statement of Theorem 10.4 on pages 203–204 of the Schaum's Outline Series book on differential geometry). On the other hand, the existence proof requires the solution of a system of partial differential equations.

In order to prove the existence of a regular smooth surface parametrization it is necessary to solve partial differential equations of the form $D\,y = A(\,x, y\,)$ where x and y are vectors and A is a smooth matrix valued function of x and y. In contrast to the situation for ordinary differential equations, the partial differential equation given above does not necessarily have a solution; specifically, the standard mixed partial derivative identities

$$\frac{\partial^{\,2}}{\partial x_i\,\partial x_j} \;=\; \frac{\partial^{\,2}}{\partial x_j\,\partial x_i}$$

imply that the entries of $A(x, y)$ and their partial derivatives must satisfy certain equations. However, the following result of F. G. Frobenius ensures that solutions always exist provided these conditions are satisfied:

FROBENIUS INTEGRABILITY THEOREM. *Let $n = k + d$, identify $\mathbf{R}^n$ with $\mathbf{R}^k \times \mathbf{R}^d$, let U be a connected domain in $\mathbf{R}^n$ and Let $\mathbf{A}$ be a smooth function defined on U and taking values in the space of $d \times k$ matrices, and let $(\mathbf{a}.\,\mathbf{b}\,) \in U$. Denote the entries of $\mathbf{A}$ by $A_{i,j}$.*

Assume in addition that these functions satisfy the compatibility conditions

$$\frac{\partial A_{i,j}}{\partial x_r} \;+\; \sum_{s=1}^{d} \;\frac{\partial A_{i,j}}{\partial x_s}\,A_{s,r} \;\;=\;\; \frac{\partial A_{i,r}}{\partial x_j} \;+\; \sum_{s=1}^{d} \;\frac{\partial A_{i,r}}{\partial x_s}\,A_{s,j}\;.$$

Then there exists a unique function Φ defined on an open disk V containing $\mathbf{a}$ and taking values in $\mathbf{R}^d$ such that the following conditions hold:

[1] $\Phi(\mathbf{a}) = \mathbf{b}$

[2] $(\,\mathbf{x},\,\Phi(\mathbf{x})\,) \in U$ *for all* $\mathbf{x} \in V$.

[3] $D\,\Phi(\mathbf{x}) = \mathbf{A}(\,\mathbf{x},\,\Phi(\mathbf{x})\,)$

Conversely, if such a function exists then the compatibility condition is satisfied.

Biographical information on Frobenius, and also many other mathematicians, may be found at the following online site:

`http://www-gap.cds.st-and.ac.uk/~history/BiogIndex.html`

The proof of the existence portion of the Fundamental Theorem of Local Surface Theory is discussed on pages 311-314 of do Carmo as well as in Appendix 2 on pages 264–265 of the Schaum's Outline Series book on differential geometry.∎

A generalization of the Fundamental Theorem of Local Surface Theory to hypersurfaces of dimension $(n-1)$ in $\mathbf{R}^n$ is established in Section 9.2 of Hicks, *Notes on Differential Geometry*; the argument is a direct generalization of the proof for surfaces.∎