Mathematics 433/533; Class Notes

Richard Koch

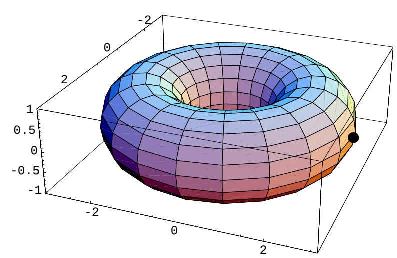
March 24, 2005

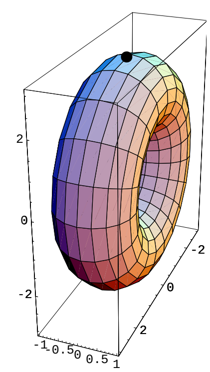
Contents

Preface

Differential geometry began in 1827 with a paper of Gauss titled *General Investigations of Curved Surfaces*. In the paper, Gauss recalled Euler's definition of the curvature of such a surface at a point p. Then he completely transformed the subject by asking a profound question.

Here is Euler's definition of curvature. Suppose p is a point on a surface $\mathcal{S}$. We can translate and rotate this surface without changing the curvature at p. Let us translate and rotate until p lies at the origin in the xy-plane and its tangent plane is this xy-plane.





After this motion, we can imagine that the new surface is given by $z = f(x, y)$. Expanding in a Taylor series about the origin, we have

$$z = f(0,0) + \left(\frac{\partial f}{\partial x} x + \frac{\partial f}{\partial y} y \right) + \frac{1}{2!} \left(\frac{\partial^2 f}{\partial x^2} x^2 + 2 \frac{\partial^2 f}{\partial x \partial y} xy + \frac{\partial^2 f}{\partial y^2} y^2 \right) + \dots$$

where all partial derivatives are computed at the origin. But since the tangent plane at the origin is the xy-plane, both partials $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ vanish and the Taylor expansion reduces to

$$f(x, y) = \frac{1}{2!} \left(\frac{\partial^2 f}{\partial x^2} x^2 + 2 \frac{\partial^2 f}{\partial x \partial y} xy + \frac{\partial^2 f}{\partial y^2} y^2 \right) + \dots$$

It is convenient to call these three second partials A, B, and C and thus write

$$f(x,y) = \frac{1}{2!}\left(Ax^2 + 2Bxy + Cy^2\right) + \dots$$

We are still free to rotate the surface about the z-axis, since such rotation leaves the tangent plane unchanged. A famous algebraic result asserts that we can eliminate the xy quadratic term by rotating the function appropriately. Many people are familiar with this result in a different context. The solution of the equation $Ax^2 + 2Bxy + Cy^2 = D$ is an ellipse in general position. If we rotate this ellipse until it is parallel to the x and y axes, its equation changes to the form $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.

Imagine that we have eliminated the xy term in our formula. The resulting numbers A and C are then called κ_1 and κ_2 in the literature, and the surface is given by the Taylor expansion

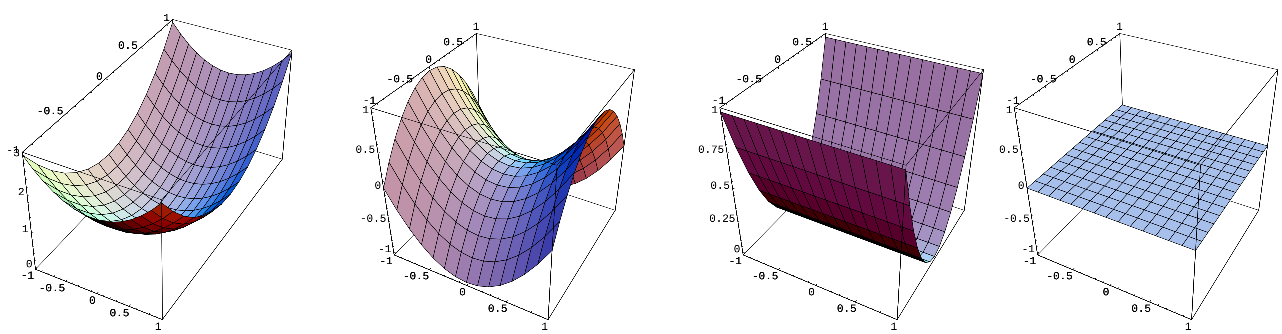
$$f(x,y) = \frac{\kappa_1}{2}x^2 + \frac{\kappa_2}{2}y^2 + \dots$$

All of this was done by Euler. The numbers κ_1 and κ_2 describe the curvature of the surface at p. Euler could compute these numbers directly from the formula $z = F(x,y)$ of the original surface without physically rotating and translating.

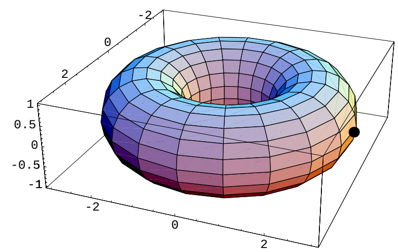
Notice that the numbers κ_1 and κ_2 depend on p and thus become functions on the surface. They are called the *principal curvatures* of the surface. The pictures below show the surfaces

$$z = \frac{\kappa_1}{2}x^2 + \frac{\kappa_2}{2}y^2$$

for both κ's positive, for one positive and one negative, for one positive and one zero, and for both zero.



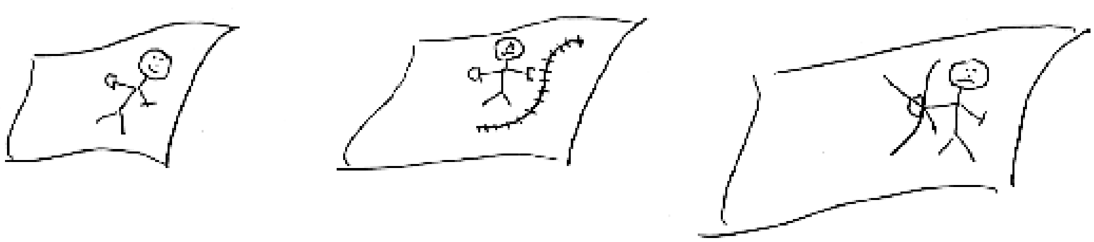
Consider the special case of a torus, shown on the next page. Along the outside rim, we will have $\kappa_1 > 0$ and $\kappa_2 > 0$ and the surface looks like a paraboloid, while along the inside rim we have $\kappa_1 > 0$ and $\kappa_2 < 0$ and the surface looks like a saddle.



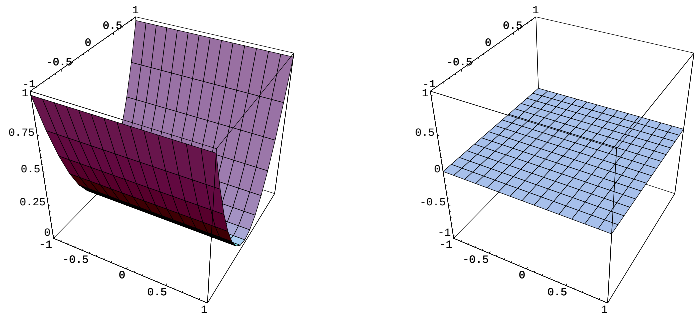
We now come to Gauss' profound question:

Can a two-dimensional person living on the surface determine κ_1 and κ_2, if that person is unable to see into the third dimension?

Let us provide our two-dimensional person with tools: an infinitesimal ruler and an infinitesimal protractor. Using the ruler, our worker can determine the lengths of curves on the surface. Rigorously, this means that our worker can integrate to determine the length of curves. Using the protractor, our worker can determine the angle between two curves. Rigorously this means that our worker can compute the angle between the tangent vectors to any two curves. Can κ_1 and κ_2 be computed with this information?



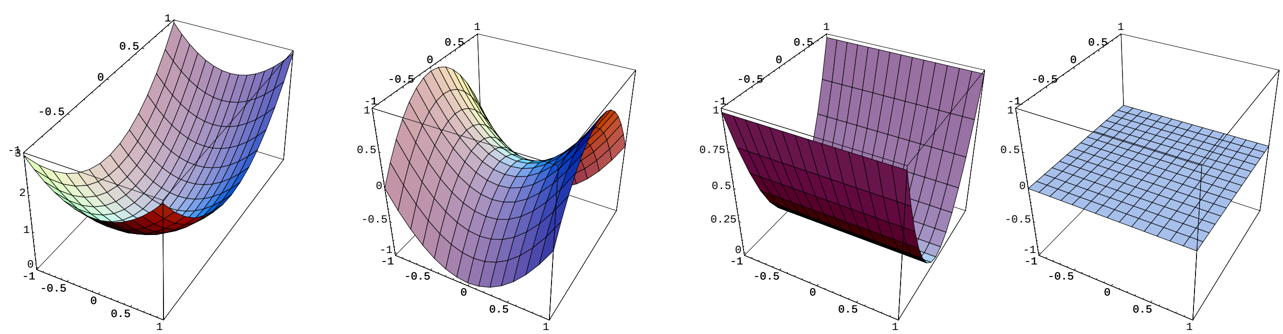
The surprising answer to Gauss' question is *no!* To see this, consider the surfaces $f(x,y) = \frac{1}{2}x^2$ and $g(x,y) = 0$, shown on the next page. For the first we have $\kappa_1 = 1$ and $\kappa_2 = 0$, while for the second we have $\kappa_1 = \kappa_2 = 0$. However, it is possible to bend the first surface until it is flat without changing the lengths of curves in the surface or the angles of curves at intersection points. So a two-dimensional worker could not tell the difference between these surfaces.



But Gauss did not give up. Before Gauss, mathematicians often wrote in an expansive way. Numerical experiments were mixed with conjectures and half-proved theorems. Gauss introduced a more austere style, which has been adopted by mathematicians ever since. He calmly stated theorems and proofs, letting his results speak for themselves. But when Gauss came to the main result in his paper on surfaces, he allowed himself to write "a remarkable theorem." And the theorem has been known as the *theorema egregium* ever since:

Theorem Egregium A two-dimensional person *can* compute the product $\kappa_1 \cdot \kappa_2$.

The number $\kappa = \kappa_1 \cdot \kappa_2$ is called the *Gaussian curvature* of the surface at p. Consider the four surfaces below again. The Gaussian curvature is positive in the first case, negative in the second case, and zero in the third and fourth cases. So a two-dimensional person can distinguish between the first two surfaces, but not between the last two.



Gauss gave several proofs of the theorema egregium. The first is computational, but the computation is very interesting. In ordinary Euclidean geometry, the distance ds between two infinitesimally close points $p = (p_1, \ldots, p_n)$ and $p + \Delta p = (p_1 + dx_1, \ldots, p_n + dx_n)$ is given by the Pythagorian theorem

$$ds^2 = dx_1^2 + \ldots + dx_n^2.$$

Gauss introduced curved coordinates (x_1, x_2) on a surface. The Pythagorian theorem is no longer true because the coordinates curve, but Gauss showed that for small dx_i it can be replaced by a formula

$$ds^2 = \sum g_{ij} dx_i dx_j$$

The g_{ij} are exactly the quantities which can be measured by our two-dimensional worker. Thus our worker can determine geometrical quantities exactly when they can be expressed in terms of the g_{ij}. Gauss found formulas for several important quantities in terms of the g_{ij} and ultimately found a formula for κ in terms of g_{ij} and their first and second derivatives. Thus a two-dimensional worker can compute κ.

I'd like to give the flavor of these computations without giving details. Since a two-dimensional worker can find the lengths of curves, the worker can determine the *shortest* curves between two points. Such curves are called *geodesics*. If $\gamma(t) = (\gamma_1(t), \dots, \gamma_n(t))$ is such a geodesic, Gauss proved that γ solves a differential equation

$$\frac{d^2\gamma_i}{dt^2} + \sum_{j,k} \Gamma^i_{jk} \frac{d\gamma_j}{dt} \frac{d\gamma_k}{dt} = 0$$

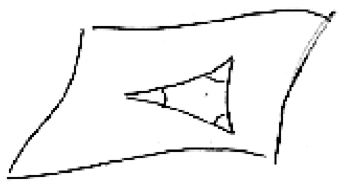
where the Γ^i_{jk}, known as the Christoffel symbols, are given in terms of the g_{ij}. Indeed,

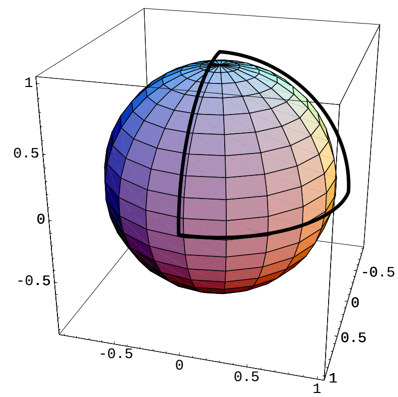
$$\Gamma^i_{jk} = \frac{1}{2} \sum_l g^{-1}_{il} \left(\frac{\partial g_{lk}}{\partial x_j} + \frac{\partial g_{lj}}{\partial x_k} - \frac{\partial g_{jk}}{\partial x_l} \right)$$

We will deduce these equations later in the course. These same equations occur in Einstein's general relativity theory, where now $\gamma(t)$ is the curve traversed by a particle acted on by gravitation, and g_{ij} describes the gravitational field.

Gauss' formula for κ is a fairly simple expression involved Γ^i_{jk} and their derivatives.

However, Gauss later found a more conceptional proof of the theorema egregium. Draw a triangle on the surface whose sides are geodesics, as illustrated on the next page.





Let α, β, γ be the three angles of this triangle. In Euclidean geometry we have $\alpha+\beta+\gamma = \pi$. This result is no longer true for curved surfaces. For example, consider the triangle illustrated above on the surface of a sphere. All three angles of this triangle equal $\frac{\pi}{2}$.

The quantity $\alpha+\beta+\gamma-\pi$ measures the discrepency between the angles of our curved triangle and the expected sum. Gauss proved the following result, now known as the theorem of Gauss-Bonnet:

Theorem of Gauss-Bonnet

$$\alpha+\beta+\gamma-\pi = \int\int_{\text{triangle}} \kappa$$

Suppose we are two-dimensional people and wish to determine κ. Lay out a triangle with geodesic sides. The number κ is a function, but assume the triangle is so small that κ is essentially constant on it. Then

$$\kappa = \frac{\alpha+\beta+\gamma-\pi}{\text{area of triangle}}$$

Example: On the spherical triangle illustrated above, all three angles are $\frac{\pi}{2}$ and the area of the triangle is one eighth the area of a sphere, so $\frac{4\pi}{8} = \frac{\pi}{2}$. Thus $\kappa = \frac{3\frac{\pi}{2}-\pi}{\pi/2} = 1$. This is correct because on a sphere of radius one we have $\kappa_1 = \kappa_2 = 1$.

One of Gauss' unusual jobs was to head a geodesic survey of Germany. During the survey, he instructed workers to climb three mountains and measure the angles between them very accurately. I'm sorry to report that the sum was π to within experimental error.

In letters to his closest friends, Gauss admitted that he was interested in the three dimensional case of his theorem. Suppose we are three-dimensional people, living in a world which curves into the fourth dimension. Can we determine that the world is curved without stepping into the fourth dimension? Gauss warned his friends not to reveal this thought, for fear that he would be thought crazy.

Chapter 1

Curves

1.1 Introduction

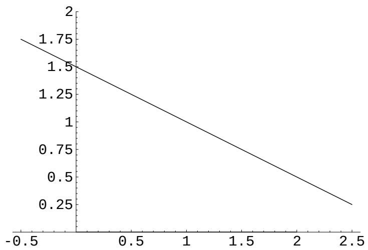
A *parameterized curve* in R^3 is a map $\gamma(t)$ from an open interval $I \subseteq R$ to R^3. In coordinates $\gamma(t) = (x(t), y(t), z(t))$. We sometimes talk of parameterized curves in R^n for arbitrary n.

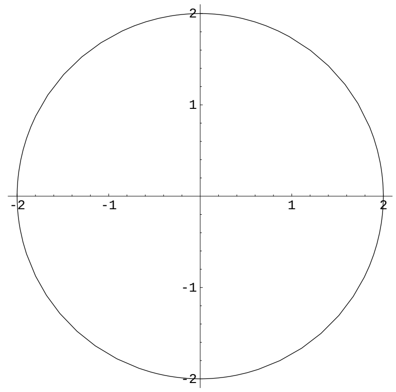
Definition 1 *We say that* $\gamma(t)$ *is a* regular curve *if the coordinate functions* $x(t), y(t), z(t)$ *are* C^∞ *and if* $\gamma'(t) = \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}\right)$ *is never zero.*

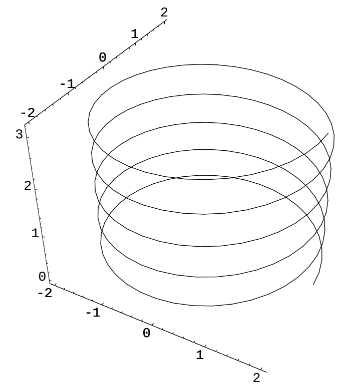
Example 1: The straight line through p in the direction v can be written $\gamma(t) = p + tv$ and thus is a parameterized curve whenever $v \neq 0$. A picture of the line $\gamma(t) = (1 + 2t, 1 - t) = (1, 1) + t(2, -1)$ is shown on the next page.

Example 2: The circle about the origin of radius r can be written $\gamma(t) = (r \cos t, r \sin t)$.

Example 3: The helix at the top of the next page can be written $\gamma(t) = (r \cos t, r \sin t, at)$ for constants r and a.

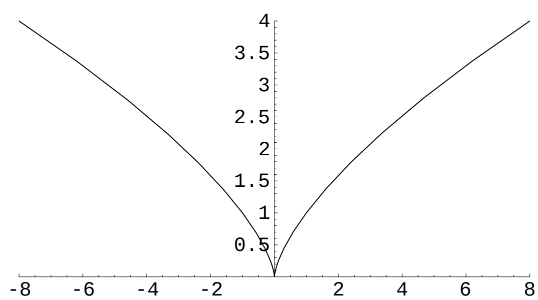






Remark: Recall that a function $f(t)$ is C^∞ if it has derivatives of all orders. Most common functions have this form. For instance, $f(t) = t^3 + 3t + 5$ is C^∞ and after a while all of its derivatives are zero. The function $f(t) = \arctan\sqrt{\frac{x+1}{x-3}}$ is C^∞ on its domain, although it would not be pleasant to write down formulas for higher derivatives. We require that the coordinate functions be C^∞ because we intend to differentiate these functions many times.

Remark: Consider the curve $\gamma(t) = (t^3, t^2)$, whose graph is shown below. This curve has a kink at the origin even though its coordinate functions are C^∞. However, $\gamma'(0) = (0,0)$. We require that $\gamma'(t)$ be nonzero to eliminate this sort of example.



1.2 Arc Length

Definition 2 *The length of $\gamma(t)$ on the interval $a \le t \le b$ is defined to be*

$$\int_a^b ||\gamma'(t)||\; dt = \int_a^b \sqrt{\left(\frac{dx_1}{dt}\right)^2 + \ldots + \left(\frac{dx_n}{dt}\right)^2}\; dt$$

Remark: In this definition, $||\gamma'(t)||$ is the length of the derivative. It is easy to make the definition plausible. Replacing the integral with a Riemann sum, we have that the length is approximately

$$\sum \left|\left|\frac{\gamma(t_i + \Delta t) - \gamma(t_i)}{\Delta t}\right|\right| \Delta t = \sum ||\gamma(t_i + \Delta t) - \gamma(t_i)||$$

which equals the length of a polyhedral approximation to the curve; in the limit we get the length of the curve.

Remark: Sadly, it is usually difficult to compute curve lengths. The circle of radius R is given by $\gamma(t) = (R\cos t, R\sin t)$ and so $||\gamma'(t)|| = ||(-R\sin t, R\cos t)|| = R$. Thus we obtain the satisfying result below for length:

$$\int_0^{2\pi} R\ dt = 2\pi R.$$

But the integral for the length of an ellipse is already beyond elementary calculation (yielding an *elliptic integral*), and the length of the parabola $y = x^2$ for $0 \le x \le b$ is just barely computable:

$$\int_0^b \sqrt{1^2 + (2x)^2}\ dx = \frac{b\sqrt{1+4b^2}}{2} + \ln\left(2b + \sqrt{1+4b^2}\right)$$

1.3 Reparameterization

Undeterred by the practical impossible of computing arc length, we now insist that curves be traversed at unit speed, so that the length of the curve for $0 \le t \le s$ is precisely s. We do this because we are interested only in the geometry of our curves, and not in how they are traversed in time.

Suppose $\gamma(t) : I \to R^3$ is a regular curve. Fix t_0 and define

$$s(t) = \int_{t_0}^t \left|\left|\gamma'(t)\right|\right|\ dt$$

Thus $s(t)$ is the length of the curve from t_0 to t and is defined on the open interval I where γ is defined.

Theorem 1 *The image $s(I)$ is an open interval and $s : I \to s(I)$ is one-to-one and onto. The map s is C^∞ and its inverse map $t = \varphi(s)$ is also C^∞.*

Proof: By the fundamental theorem of calculus, $\frac{ds}{dt} = ||\gamma'(t)||$; since γ is a regular curve, $\gamma'(t)$ is not zero, so this derivative is positive. Hence s is a strictly increasing function and thus one-to-one.

Since s is continuous, the image $s(I)$ is a connected subset of R and thus an interval (possibly infinite). The interval is open, because if $s_1 = s(t_1)$ is in the image, then we can find $a < t_1 < b$ in the open interval I and so $s(a) < s(t_1) = s_1 < s(b)$ in the image interval.

Finally we will prove that $\varphi(s)$ is differentiable. By definition, the derivative of φ at s_1 is

$$\lim_{s \to s_1} \frac{\varphi(s) - \varphi(s_1)}{s - s_1}$$

Notice that we have abused the notation by letting the letter 's' stand for the function $s(t)$ and also an arbitrary point in the image interval $s(I)$. We will continue this abuse (!!) by writing $s = s(\varphi(s))$ and $s_1 = s(\varphi(s_1))$ so the limit becomes

$$\frac{d\varphi}{ds} = \lim_{s \to s_1} \frac{1}{\frac{s(\varphi(s)) - s(\varphi(s_1))}{\varphi(s) - \varphi(s_1)}}$$

However, $\varphi(s)$ is some number t and $\varphi(s_1)$ is a number t_1 and the denominator thus equals $\lim_{t \to t_1} \frac{s(t)-s(t_1)}{t-t_1} = \frac{ds}{dt} = ||\gamma'(t_1)||$. So

$$\left.\frac{d\varphi}{ds}\right|_{s_1} = \left.\frac{1}{||\gamma'(t)||}\right|_{t_1} = \frac{1}{||\gamma'(\varphi(s_1))||}$$

and in particular this derivative exists. (Here we have assumed the subtle point that $s \to s_1$ implies $t \to t_1$. We leave the verification to the reader.)

It follows rapidly that $\varphi(s)$ is C^∞. Indeed $\gamma(t)$ is C^∞ and we have just proved that φ is differentiable, so $\frac{1}{||\gamma'(\varphi(s))||}$ is differentiable. But this equals $\frac{d\varphi}{ds}$, so this function is differentiable and φ is C^2. We can continue to bootstrap in this manner and show that φ has derivatives of all orders. QED.

Definition 3 *Suppose $\gamma(t)$ is a regular curve. Fix t_0 and define $s(t)$ and $\varphi(s)$ as above. Then $\gamma(\varphi(s))$ is called* the reparameterization of γ by arc-length.

Remark: Notice that the new curve is defined on an interval which contains zero (because $s(t_0) = 0$). The derivative vector of the new curve has length one because

$$\frac{d}{ds}\gamma(\varphi(s)) = \gamma'(\varphi(s))\frac{d\varphi}{ds} = \gamma'(\varphi(s))\frac{1}{||\gamma'(\varphi(s))||},$$

which is a vector of length one. It follows immediately that the length of the new curve from $s = 0$ to s is exactly s.

We now introduce a final abuse of notation. Rather than writing $\gamma(\varphi(s))$, we shall write $\gamma(s)$ for the curve parameterized by arclength. We are committing a fairly serious sin, because we do not get from $\gamma(t)$ to $\gamma(s)$ by changing the letter t to s, as is the convention in all other mathematics. Instead, we compute $s(t)$, find its inverse $\varphi(s)$, and substitute $\varphi(s)$ for t in the original formula.

In the rest of the course, $\gamma(t)$ denotes an arbitrary regular curve and $\gamma(s)$ denotes a curve parameterized by arclength.

Example: Consider the helix $\gamma(t) = (r\cos t, r\sin t, at)$. Then $\gamma'(t) = (-r\sin t, r\cos t, a)$ and $||\gamma'(t)|| = \sqrt{r^2+a^2}$. So $s(t) = \int_0^t \sqrt{r^2+a^2}\, dt = t\sqrt{r^2+a^2}$. We obtain the inverse $\varphi(s)$ by solving for t in terms of s:

$$\varphi(s) = \frac{s}{\sqrt{r^2+a^2}}$$

So

$$\gamma(s) = \left(r\cos\frac{s}{\sqrt{r^2+a^2}}, r\sin\frac{s}{\sqrt{r^2+a^2}}, \frac{as}{\sqrt{r^2+a^2}} \right)$$

1.4 Curvature

If $\gamma(s)$ is a curve parameterized by arclength, we define $T(s) = \gamma'(s)$. Notice that T is a unit vector. We call it the *tangent vector.*

Theorem 2 *The derivative* $\frac{dT}{ds}$ *is perpendicular to* $T(s)$*.*

Proof: Since $T(s)\cdot T(s) = 1$, we have

$$\frac{d}{ds}\left(T(s)\cdot T(s)\right) = \frac{dT}{ds}\cdot T(s) + T(s)\cdot\frac{dT}{ds} = 0$$

and so $\frac{dT}{ds}\cdot T = 0$. QED.

Definition 4 *If* $\gamma(s)$ *is parameterized by arclength, we define the curvature* $\kappa(s)$ *by*

$$\kappa(s) = \left|\left|\frac{dT}{ds}\right|\right|$$

If this curvature is not zero, we define the normal vector and binormal vector by

$$N(s) = \frac{\frac{dT}{ds}}{||\frac{dT}{ds}||}$$

$$B(s) = T(s)\times N(s)$$

The main point of this section is to convince you that $\kappa(s)$ is a reasonable measure of the *curvature* of $\gamma(s)$. It is easy to see that κ is a rough measure of curvature: if our curve always goes in the same direction, $T(s)$ points constantly in this direction and its derivative is zero. So when $\kappa(s) = ||\frac{dT}{ds}||$ is not zero, the curve must be changing direction.

However, we will prove something more precise. Approximate γ near s by a circle. Classically this circle was called the *osculating circle* or *kissing circle*. We will prove that κ equals one over the radius of this circle. If γ is curving rapidly, then R is small and so κ is large, as we'd wish. If γ is curving very slowly, then R is enormous and so κ is small, as we'd wish. If γ is a straight line, then $R = \infty$ and so $\kappa = 0$.

Theorem 3 *The curve $\gamma(s)$ is a straight line if and only if $\kappa(s)$ is always zero.*

Proof: If $\gamma(s) = \vec{p} + s\vec{v}$, then $T(s) = \vec{v}$, a constant vector, so $\frac{dT}{ds} = 0$. Conversely, if $\kappa = 0$ then $\frac{dT}{ds} = 0$ and thus T is a constant vector $\vec{v}$. Since $T = \gamma'(s)$, we can integrate to get $\gamma(s) = \vec{p} + s\vec{v}$. QED.

Theorem 4 *Let $\gamma(s)$ be a curve parameterized by arclength and fix a point s_0. There is a unique circle $C(s)$ in R^3 parameterized by arclength such that C and γ agree at s_0 up to derivatives of order two, so $C(s_0) = \gamma(s_0), C'(s_0) = \gamma'(s_0), C''(s_0) = \gamma''(s_0)$. If this circle has radius R, then*

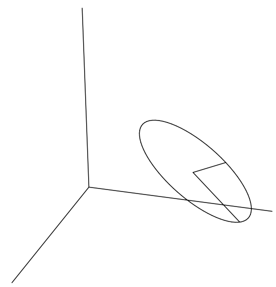
$$\kappa(s_0) = \frac{1}{R}.$$

Proof: First we develop a formula for a parameterized circle in general position in R^3. Let $\vec{p}$ be the center of the circle. Let $\vec{v}_1$ be a unit vector pointing from p to $\gamma(s_0)$ and let $\vec{v}_2$ be a unit vector perpendicular to $\vec{v}_1$. Then $\vec{v}_1$ and $\vec{v}_2$ define a plane and the circle of radius R in this plane can be parameterized by

$$C(t) = \vec{p} + (R\cos t)\vec{v}_1 + (R\sin t)\vec{v}_2.$$

Let us replace t by $\frac{s-s_0}{R}$ so C will be parameterized by arclength and be at the correct location when $s = s_0$. Then

$$C(s) = \vec{p} + (R\cos\frac{s-s_0}{R})\vec{v}_1 + (R\sin\frac{s-s_0}{R})\vec{v}_2.$$



We now try to match $C(s)$ and $\gamma(s)$. We have

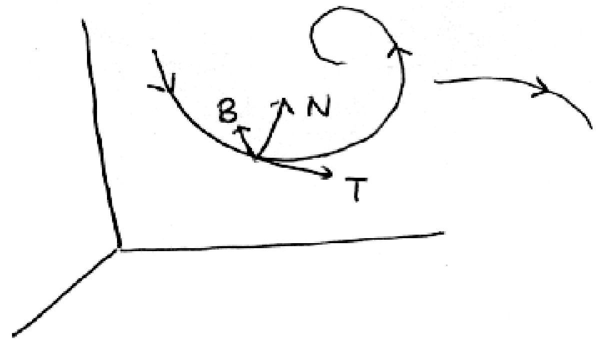
$$\begin{aligned} C(s_0) &= \vec{p} + R\vec{v}_1 = \gamma(s_0) \\ C'(s_0) &= \vec{v}_2 = \gamma'(s_0) = T(s_0) \\ C''(s_0) &= -\frac{1}{R}\vec{v}_1 = \gamma''(s_0) = \kappa(s_0)N(s_0) \end{aligned}$$

These equations have a unique solution. Therefore the osculating circle is completely determined and in particular $\kappa = \frac{1}{R}$.

$$\begin{aligned} R &= \frac{1}{\kappa} \\ \vec{v}_1 &= -N \\ \vec{v}_2 &= T \\ \vec{p} &= \gamma(s_0) + \frac{1}{\kappa}N \end{aligned}$$

1.5 The Moving Frame

In this section we shall suppose that $\gamma(s)$ is parameterized by arc length and $\kappa(s)$ is never zero. In this situation, we have attached three orthonormal vectors $T(s), N(s), B(s)$ to each point of the curve. These three vectors form a basis of R^3, which moves along the curve. Geometers use the word *frame* instead of the word basis, so T, N, B is called the moving frame.



Imagine an isolated point traveling along the curve at constant speed. The moving frame allows us to replace this single point with an airplane. The nose of the airplane should point along T, the left wing along N, and the tail along B. As we travel along the curve, the airplane pitches and rolls as T, N, and B move. At each point, the orientation of the plane is completely determined by T, N, and B.

As we'll see, the introduction of T, N, and B is decisive in the theory.

1.6 The Frenet-Serret Formulas

We want to measure the change of the moving frame as we travel along the curve. So we differentiate all three vectors. Since T, N, and B form a basis, we can write these derivatives as linear combinations of T, N, and B, and it turns it to be very important to do this instead of expressing them as linear combinations of the standard basis.

Let us temporarily define $X_1 = T, X_2 = N, X_3 = B$. Then the expression of the derivatives of the frame vectors as a linear combination of these vectors is given by coefficients a_{ij} such that

$$\frac{dX_i}{ds} = \sum_j a_{ij} X_j$$

Theorem 5 *The matrix $A = (a_{ij})$ is skew symmetric, so $a_{ji} = -a_{ij}$. In particular, $a_{ii} = 0$.*

Proof: Since the basis T, N, B is orthonormal, we have $X_i \cdot X_j = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. Differentiating with respect to s yields $\frac{dX_i}{ds} \cdot X_j + X_i \cdot \frac{dX_j}{ds} = \frac{d}{ds}\delta_{ij} = 0$. By the previous displayed formula, this can be written

$$\left(\sum_k a_{ik} X_k \right) \cdot X_j + X_i \cdot \left(\sum_k a_{jk} X_k \right)$$

and so

$$\sum_k a_{ik} X_k \cdot X_j + \sum_k a_{jk} X_i \cdot X_k = 0$$

But $X_k \cdot X_j$ is zero unless $k = j$, so this equation reduces to

$$a_{ij} + a_{ji} = 0.$$

QED

Returning to the notation T, N, B, the previous theorem allows us to write

$$\begin{array}{rcccccc} \frac{dT}{ds} & = & a_{11}T & + & a_{12}N & + & a_{13}B \\ \frac{dN}{ds} & = & a_{21}T & + & a_{22}N & + & a_{23}B \\ \frac{dB}{ds} & = & a_{31}T & + & a_{32}N & + & a_{33}B \end{array}$$

Applying the theorem, these equations simplify to

$$\begin{array}{rcccccc} \frac{dT}{ds} & = & & & a_{12}N & + & a_{13}B \\ \frac{dN}{ds} & = & -a_{12}T & & & + & a_{23}B \\ \frac{dB}{ds} & = & -a_{13}T & & -a_{23}N & & \end{array}$$

However, we already know that $\frac{dT}{dt} = \kappa N$, so the equations simply further to

$$\begin{array}{rcccccc} \frac{dT}{ds} & = & & & \kappa N & & \\ \frac{dN}{ds} & = & -\kappa T & & & + & a_{23}B \\ \frac{dB}{ds} & = & & & -a_{23}N & & \end{array}$$

Definition 5 *The quantity a_{23} is called the* torsion *of the curve and written $\tau(s)$. It is a function of s.*

We have proved the Frenet-Serret formulas, first discovered in 1851. We'll soon see that these formulas hold the key to the theory of curves:

Theorem 6 (Frenet-Serret) *Let $\gamma(s)$ be a curve parameterized by arc length and suppose the curve $\kappa(s)$ is never zero. Then*

$$\begin{aligned} \frac{dT}{ds} &= \kappa(s)\ N \\ \frac{dN}{ds} &= -\kappa(s)\ T + \tau(s)\ B \\ \frac{dB}{ds} &= -\tau(s)\ N \end{aligned}$$

We end this section with a description of the geometrical meaning of torsion. Fix s and recall that our curve is moving in the direction $T(s)$ and is approximated by a circle with center in the direction $N(s)$. These two vectors define a plane in which the curve is momentarily trapped. Classical geometers called this plane the *osculating plane*, which means *kissing plane*. Since $B(s)$ is perpendicular to this plane, it forms a handle for the osculating plane exactly like the handle on the tool used by carpenters to smooth plaster walls.

If the curve remains in this osculating plane forever, then $B(s)$ should be constant and its derivative $-\tau(s)N(s)$ should be zero, so $\tau = 0$. Otherwise, τ measures the curve's *attempt to twist out of the plane in which it finds itself momentarily trapped.*

Let us convert this informal discussion into a theorem:

Theorem 7 *A curve $\gamma(s)$ lies in a plane if and only if $\tau(s)$ is identically zero.*

Proof: Suppose γ lies in a plane. Then

$$T(s) = \lim_{h\to 0} \frac{\gamma(s+h) - \gamma(s)}{h}$$

is parallel to the plane, and consequently

$$\kappa N = \lim_{h\to 0} \frac{T(s+h) - T(s)}{h}$$

is also parallel to the plane. It follows that $B = T \times N$ is perpendicular to the plane for all s. Since B has unit length and varies continuously with s, it must be constant, so its derivative $-\tau(s)\ N$ is zero.

Conversely, suppose $\tau = 0$. Then B is constant. Fix s_0 and consider the plane

$$\mathcal{P} = \{\ x \mid (x - \gamma(s_0)) \cdot B = 0\ \}$$

We claim that γ is in this plane. To see this, substitute $\gamma(s)$ for x. The expression $(\gamma(s) - \gamma(s_0)) \cdot B$ is constant because its derivative is $T \cdot B = 0$. When $s = s_0$, this constant is zero, so it is always zero and $\gamma(s)$ always satisfies the equation of the plane. QED.

Example: Finally, we will compute the curvature and torsion of a helix. We earlier discovered that the equation of a helix parameterized by arc length is

$$\gamma(s) = \left(r\cos\frac{s}{\sqrt{r^2+a^2}}, r\sin\frac{s}{\sqrt{r^2+a^2}}, \frac{as}{\sqrt{r^2+a^2}} \right)$$

For this helix,

$$T(s) = \frac{1}{\sqrt{r^2+a^2}} \left(-r\sin\frac{s}{\sqrt{r^2+a^2}}, r\cos\frac{s}{\sqrt{r^2+a^2}}, a \right)$$

and so

$$\frac{dT}{ds} = \frac{1}{r^2+a^2} \left(-r\cos\frac{s}{\sqrt{r^2+a^2}}, -r\sin\frac{s}{\sqrt{r^2+a^2}}, 0 \right)$$

The length of this derivative is the curvature, so

$$\kappa(s) = \frac{r}{r^2+a^2}$$

The normal vector is the unit vector in the direction of $\frac{dT}{ds}$, so

$$N(s) = \left(-\cos\frac{s}{\sqrt{r^2+a^2}}, -\sin\frac{s}{\sqrt{r^2+a^2}}, 0 \right)$$

Then $B(s) = T(s) \times N(s)$ and this expression equals $\frac{1}{\sqrt{r^2+a^2}}$ times

$$\left(-r\sin\frac{s}{\sqrt{r^2+a^2}}, r\cos\frac{s}{\sqrt{r^2+a^2}}, a \right) \times \left(-\cos\frac{s}{\sqrt{r^2+a^2}}, -\sin\frac{s}{\sqrt{r^2+a^2}}, 0 \right)$$

or

$$B(s) = \frac{1}{\sqrt{r^2+a^2}} \left(a\sin\frac{s}{\sqrt{r^2+a^2}}, -a\cos\frac{s}{\sqrt{r^2+a^2}}, r \right)$$

It is supposed to be true that $\frac{dB}{ds} = -\tau(s)N(s)$, and indeed

$$\frac{dB}{ds} = \frac{a}{r^2+a^2} \left(\cos\frac{s}{\sqrt{r^2+a^2}}, \sin\frac{s}{\sqrt{r^2+a^2}}, 0 \right)$$

so we have

$$\tau(s) = \frac{a}{r^2+a^2}.$$

1.7 The Goal of Curve Theory

Let's step back and develop some goals. We want to understand the *geometry* of curves. Around 1870, Felix Klein gave a famous lecture in Erlangen, Germany, defining geometry. In the lecture, Klein called attention to the key role played in geometry by Euclidean motions, that is, maps preserving distances and angles. Among these maps are translations, rotations, and reflections. According to Klein, *geometry is the study of properties of figures which are invariant under Euclidean motions.* For example, suppose we have a line segment from $(1,1)$ to $(4,5)$. Then the x coordinate of the starting point is not an interesting quantity because it changes if we translate the picture. But the length of this segment, $\sqrt{3^2+4^2}=5$, is interesting because it remains unchanged under translation and rotation of the segment.

Let us apply this philosophy to curve theory. Unlike the physicists, we are not interested in how the curve is traced in time, but only in the curve itself. And we are not interested in properties which change when the curve is picked up and moved somewhere else, but only in quantities which do not change under such motions. Clearly the curvature and torsion, κ and τ, are invariant under such motions and therefore genuine geometric quantities.

Astonishingly, they are the only geometric quantities. Every other interesting quantity can be written as an expression in curvature and torsion. A better way to say this is that κ and τ completely determine the curve up to Euclidean motions. If two curves have the same curvature and torsion functions, then one can be rotated and translated until it lies right on top of the other.

1.8 The Fundamental Theorem of Curve Theory

Theorem 8 (The Fundamental Theorem)

- *Let γ be a curve parameterized by arclength and suppose the curvature of γ never vanishes. Then the functions $\kappa(s)$ and $\tau(s)$ completely determine the curve up to Euclidean motions. That is, if $\sigma(s)$ is a second curve with the same curvature and torsion, then there is a Euclidean motion M such that $\sigma(s) = M \circ \gamma(s)$.*

- *Conversely, suppose $\kappa(s)$ and $\tau(s)$ are C^∞ functions with $\kappa(s) > 0$. Then there is a curve $\gamma(s)$ parameterized by arclength with κ and τ as curvature and torsion.*

Proof: Consider the system of differential equations below:

$$\begin{aligned}
\frac{d\gamma}{ds} &= T \\
\frac{dT}{ds} &= \kappa(s)\, N \\
\frac{dN}{ds} &= -\kappa(s)\, T + \tau(s)\, B \\
\frac{dB}{ds} &= -\tau(s)\, N
\end{aligned}$$

We'll fill in details in a moment, but the essence of the proof is easy. These equations are linear, so by the existence and uniqueness theorem for differential equations, they have a unique solution once initial values are given. These initial values are $\gamma(0)$, which is the starting point of the curve, and $T(0), N(0), B(0)$, which is the starting orientation. Thus the differential equations completely determine the curve up to a Euclidean motion.

Let's expand the argument. We'll first prove existence. Solve the equations with initial values $\gamma(0) = (0,0,0), T(0) = (1,0,0), N(0) = (0,1,0)$, and $B(0) = (0,0,1)$. We claim that $T(s), N(s), B(s)$ are orthonormal for all s. For a moment, assume this. But then $\frac{d\gamma}{ds} = T$ is always a unit vector, so $\gamma(s)$ is a curve parameterized by arc length. Moreover $\frac{dT}{ds} = \kappa N$, so κ must be the curvature and N the vector normal to the curve. Since B has unit length, is perpendicular to T and N, and equals $T \times N$ when $s = 0$, it must equal $T \times N$ always by continuity. Since $\frac{dB}{ds} = -\tau N$, τ is the torsion.

Next we prove the missing orthonormality assertion. Define $X_1 = T, X_2 = N, X_3 = B$ as in section 1.6. By the argument in that section,

$$\frac{d}{ds} X_i \cdot X_j = \left(\sum_k a_{ik} X_k \right) \cdot X_j + X_i \cdot \left(\sum_k a_{jk} X_k \right)$$

Temporarily let $L_{ij} = X_i \cdot X_j$ and notice that we obtain a system of differential equations for these functions:

$$\frac{d}{ds} L_{ij} = \sum_k a_{ik} L_{kj} + \sum_k a_{jk} L_{ik}$$

From differential equation theory, the solution of this equation is unique once initial conditions are given. From the initial choice of $T(0), N(0), B(0)$ we see that $L_{ij}(0) = \delta_{ij}$. But if we define $L_{ij}(s) = \delta_{ij}$ for all s, we obtain one solution of the differential equations since then $\frac{d}{ds}\delta_{ij} = 0$ and $\sum_k a_i k \delta_{kj} + \sum_k a_{jk} \delta_{ik} = a_{ij} + a_j i = 0$ by the Frenet-Serret formulas. So our solution $X_i \cdot X_j$ must equal this solution δ_{ij} and in particular the X_i are orthonormal.

Finally we prove uniqueness. This time, suppose that γ is given in advance. Call the curve obtained in the previous paragraph using special initial properties $\sigma(s)$. Translate γ

to make its starting point the origin instead of $\gamma(0)$. Then rotate γ so the three vectors $T(0), N(0), B(0)$ associated with γ rotate into the vectors $(1,0,0),(0,1,0),(0,0,1)$. This is possible because both collections are right handed orthonormal coordinate systems. The combination of the translation and rotation just used is a Euclidean motion M. Clearly Euclidean motions do not change curvature and torsion. So $M \circ \gamma$ and σ both satisfy the differential equations and have the same initial conditions, and therefore are equal. It follows that any γ is just σ up to a Euclidean motion. QED.

1.9 Computers and the Fundamental Theorem

In the previous section, we proved the fundamental theorem using fancy differential equation theory. But I prefer to think of the proof as a Kansas farmer would. Suppose we are given functions $\kappa(s)$ and $\tau(s)$. Pick a small interval Δs, say $\Delta s = .01$, and let $s_i = i\ \Delta s$. We will show how to compute $\gamma(s_i)$ numerically using a computer. Start by choosing initial conditions, say $\gamma(0) = (0,0,0), T(0) = (1,0,0), N(0) = (0,1,0)$, and $B(0) = (0,0,1)$.

Consider the differential equation $\frac{d\gamma}{ds} = T(s)$. Replace the derivative on the left side by a difference quotient to obtain

$$\frac{\gamma(s_i + \Delta s) - \gamma(s_i)}{\Delta s} = T(s_i)$$

and solve to obtain

$$\gamma(s_i + \Delta s) = \gamma(s_i) + \Delta s\ T(s_i)$$

Do the same thing with each of the Frenet-Serret equations to obtain the following iterative scheme:

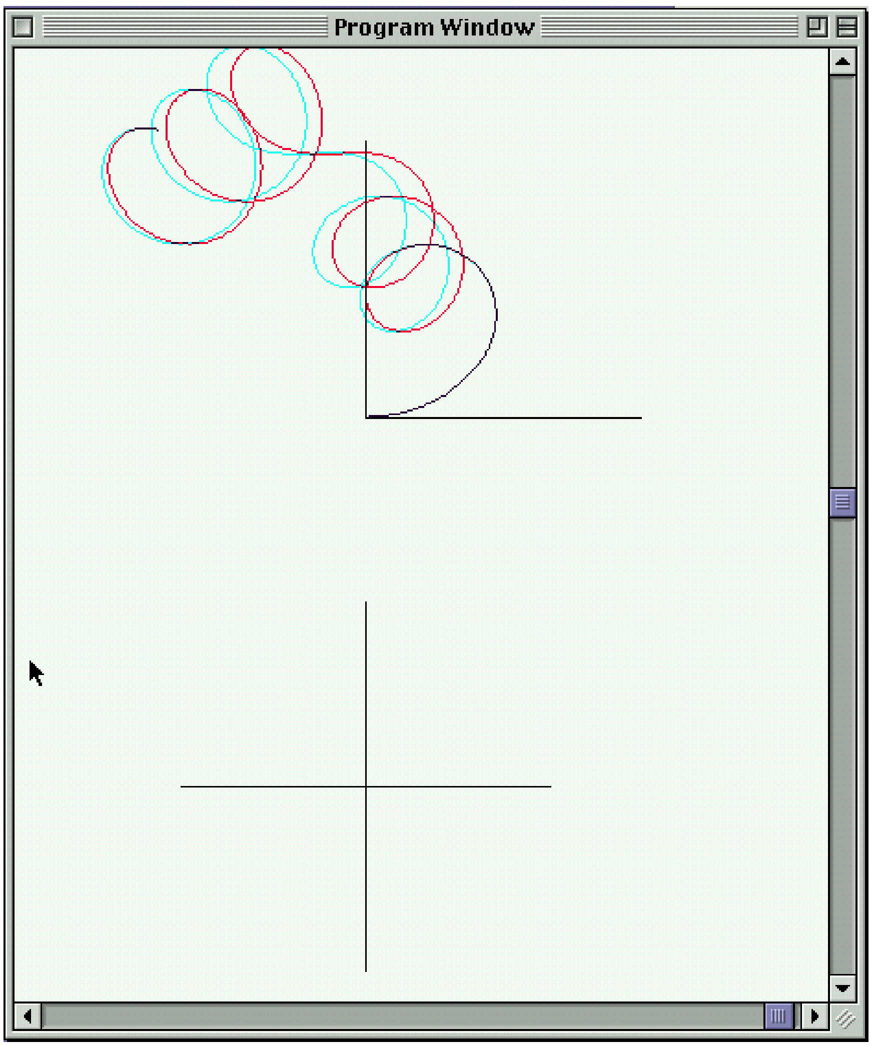
$$\left\{\begin{array}{lllllllll}
\gamma(s_i+\Delta s) & = & \gamma(s_i) & + & \Delta s & (& T(s_i) & &) \\
T(s_i+\Delta s) & = & T(s_i) & + & \Delta s & (& & \kappa(s_i)\ N(s_i) &) \\
N(s_i+\Delta s) & = & N(s_i) & + & \Delta s & (& -\kappa(s_i)\ T(s_i) & +\tau(s_i)\ B(s_i) &) \\
B(s_i+\Delta s) & = & B(s_i) & + & \Delta s & (& & -\tau(s_i)\ N(s_i) &)
\end{array}\right.$$

The initial conditions determine $\gamma(s_0), T(s_0), N(s_0), B(s_0)$ and these formulas then determine $\gamma(s_1), T(s_1), N(s_1), B(s_1)$. Applying the formulas again gives $\gamma(s_2), T(s_2), N(s_2), B(s_2)$. Etc.

This calculation is very easy on a computer. However, numerical errors lead to T, N, and B which gradually have lengths not equal to one.. To improve accuracy, renormalize at each stage to get unit vectors. To improve accuracy still more, use the Gram-Schmidt process at each stage to insure that the frame remains orthonormal.

My favorite way to do this calculation is to allow the user to enter κ and τ in real time with a mouse. Draw a two-dimensional coordinate plane at the bottom of the screen. Input κ by moving the mouse left or right. Input τ by moving the mouse up or down. While the mouse moves, let the computer draw the curve at the top of the screen.

Below is a picture of this program in action, together with a three dimensional picture of a curve in red and green. Use red and green glasses to see the curve in 3D.



1.10 The Possibility that $\kappa = 0$

There is something strange about the proof of the fundamental theorem. It seems to work fine even if κ is sometimes zero or sometimes negative. What is going on?

Notice that the fundamental theorem is really a theorem about parameterized curves with associated moving frames. We can capture this distinction with a formal definition:

Definition 6 *A* framed curve *is a C^∞ curve $\gamma(s)$ parameterized by arclength, together with a moving frame $T(s), N(s), B(s)$ of orthonormal vectors, such that $\frac{d\gamma}{ds} = T(s)$ and $\frac{dT}{ds}$ is a multiple of $N(s)$.*

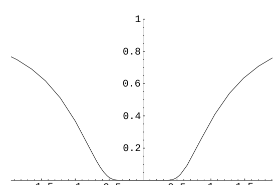
According to the fundamental theorem, such framed curves are completely determined by $\kappa(s)$ and $\tau(s)$ up to Euclidean motion, and every κ and τ can occur. Here κ is allowed to be positive, negative, or zero.

There is a natural question to ask. If $\gamma(s)$ is a curve, how many framed curves can be constructed using γ? If the answer were "exactly one," then our theory would apply without change to all $\gamma(s)$.

Example 1: We are going to show that some C^∞ curves γ have no associated framed curve at all. Start with the function

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

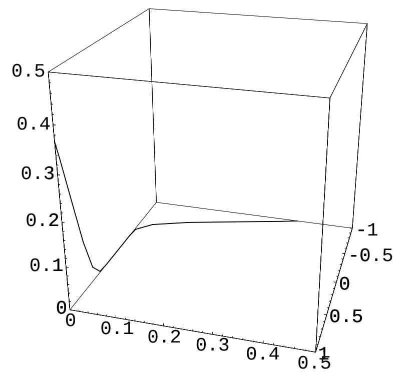
This function is pictured below. It is easy to prove that f is C^∞ and *all* derivatives at the origin are zero.



Let

$$\gamma(t) = \begin{cases} (t, f(t), 0) & \text{if } t \leq 0 \\ (t, 0, f(t)) & \text{if } t > 0 \end{cases}$$

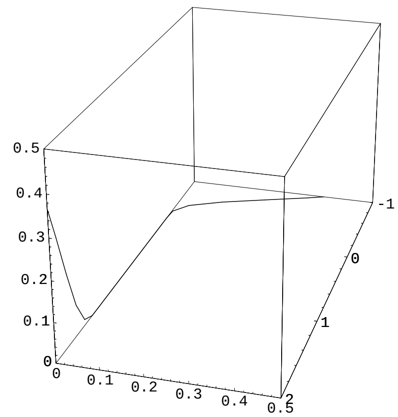
This curve is C^∞ because all derivatives of f vanish at the origin. Its derivative is never zero, so it is a regular curve and can be reparameterized by arc length. A short calculation shows that the curvature is zero only at the origin. The curve lies in the xy-plane for $t < 0$ and in the xz-plane for $t > 0$.



When $t = 0$, the tangent vector is $(1, 0, 0)$. For t just slightly smaller, the vector N must be a unit vector perpendicular to $(1, 0, 0)$ in the xy-plane and so $\pm(0, 1, 0)$. For t just slightly larger, the vector N must be a unit vector perpendicular to $(1, 0, 0)$ in the xz-plane, and so $\pm(0, 0, 1)$. Clearly there is no continuous extension of N to $t = 0$.

Example 2: Next we shall construct a C^∞ curve with infinitely many different framed extensions. We do this by gluing an interval of length one along the x-axis into the previous example. Thus let

$$\gamma(t) = \begin{cases} (t, f(t), 0) & \text{if } t \leq 0 \\ (t, 0, 0) & \text{if } 0 < t < 1 \\ (t, 0, f(t-1)) & \text{if } 1 < t \end{cases}$$



In this example, the normal vector N is $(0,1,0)$ at $t=0$ and $(0,0,1)$ at $t=2$. Somehow, it must rotate from one position to the other during the interval $0 \le t \le 1$. During this time, $T=(1,0,0)$ and B will be determined by N. It is easy to extend $N(t)$ to the interval $0 < t < 1$; we need only make N rotate in the yz-plane from one position to the other. But we can do this infinitely many ways. We can rotate once, or do many loops, or rotate positively and then negatively and then positively again.

Using the airplane example, our plane is moving straight ahead for $0 \le t \le 1$. During this time, we can perform as many rolls as we wish.

What does this example say in terms of κ and τ? Notice that during the crucial interval, $\kappa = 0$. We perform rolls by changing τ. So the conclusion is that many different combinations of κ and τ give the same curve γ.

Remark: On the other hand, during intervals when $\kappa \neq 0$, we have only two choices for frames. We can choose as in our previous theory making $\kappa > 0$, or we can change the signs of κ, N, and B. Notice that τ does not change sign because $\frac{dB}{ds} = \tau N$, so changing the signs of N and B leaves τ unchanged. So there are two framed curves for each γ and exactly one of the two has positive curvature.

Remark: The examples above show that when $\kappa \neq 0$, there are only two possible extensions of γ to a framed curve. But as soon as $\kappa = 0$, there suddenly may be no extensions, or infinitely many. However, there is an important case when these difficulties do not occur. Recall that a function is *analytic* if it can be expanded in a power series. All standard functions from calculus are analytic. If $\gamma(t)$ is analytic, it is easy to show that the renormalized curve $\gamma(s)$ is analytic.

Suppose that $\gamma(s)$ is analytic. We are going to examine the theory near a particular s_0; we may as well take $s_0 = 0$. Expand in a series

$$\gamma(s) = a_0 + a_1 s + a_2 s^2 + \ldots$$

Of course γ is a vector, so the coefficients a_i are vectors as well.

Differentiating, we find that $T(s) = a_1 + 2a_2 s + \ldots$. Differentiating again gives an expression $2a_2 + 3\cdot 2a_3 s + 4\cdot 3a_4 s^2 + \ldots$, which we would like to write as a product of two nice functions $\kappa(s)N(s)$. There are two possibilities. It may happen that all a_i for $i \geq 2$ are zero. In that case $\gamma(s)$ is a line and a little thought shows that lines can be extended to framed curves in infinitely many ways, even if we require that the extension be analytic. The other possibility is that some first a_i is not zero. In that case, write

$$\frac{dT}{ds} = b_k s^k + b_{k+1} s^{k+1} + \ldots$$

where $b_k \neq 0$.

But then $\frac{dT}{ds} = s^k\left(b_k + b_{k+1}s + \ldots\right) = s^k X(s)$ where $X(s)$ is an analytic vector which does not vanish near $s = 0$. So we can write

$$\frac{dT}{ds} = \left(s^k ||X(s)||\right) \cdot \frac{X(s)}{||X(s)||} = \left(s^k ||X(s)||\right) N(s) = \kappa(s)N(s)$$

where $N(s)$ is an analytic unit vector and $\kappa(s)$ is analytic.

Notice that κ might change sign at s. Imagine that we started the theory for γ by defining κ to be positive whenever it is not zero, and then choosing $N(s)$ as we are forced to do. The calculation we have just done proves that to extend κ and N to points where $\kappa = 0$, we may have to change the signs of κ and N on one side of the zero.

Theorem 9 *Suppose $\gamma(s)$ is analytic, and not a straight line. Then $\kappa(s)$ is zero only at isolated spots. It is possible to define κ, N, and B at all points of γ including the isolated points where $\kappa = 0$ provided there is no restriction on the sign of κ. In that case, there are exactly two ways to frame γ, and one of the two is singled out by picking the sign of κ at one single point where $\kappa \neq 0$. It is possible to go from one framing to the other by changing the signs of κ, N, and B.*

There is a one-to-one correspondence between equivalence classes of analytic curves parameterized by arc length (up to equivalence by Euclidean motion), and equivalence classes of pairs $\{\kappa(s),\ \tau(s)\}$ (up to equivalence by changing the sign of κ), except that all pairs of the form $\{0,\ \tau(s)\}$ correspond to the same curve, namely a straight line.

1.11 Formulas for $\gamma(t)$

There is an air of impracticality about our theory, because renormalization from $\gamma(t)$ to $\gamma(s)$ is usually impossible to compute explicitly, and yet we need this renormalization to compute κ and τ. We are going to remove that impracticality by producting formulas for κ and τ which do not depend on initial renormalization.

Theorem 10 *Let $\gamma(t)$ be a regular curve. Then*

$$
\begin{aligned}
T &= \frac{\gamma'(t)}{||\gamma'(t)||} \\
B &= \frac{\gamma'(t)\times\gamma''(t)}{||\gamma'(t)\times\gamma''(t)||} \\
N &= B\times T \\
\kappa &= \frac{||\gamma'\times\gamma''||}{||\gamma'||^3} \\
\tau &= \frac{(\gamma'\times\gamma'')\cdot\gamma'''}{||\gamma'\times\gamma''||^2}
\end{aligned}
$$

Proof: The vector $\gamma'(t)$ is tangent to the curve, so $T(t)$ is the normalized form $\frac{\gamma'(t)}{||\gamma'(t)||}$.

Recall that $s(t)$ is the length of the curve to t and $\frac{ds}{dt} = ||\gamma'(t)||$. Let $T(s)$ denote the usual tangent vector for the renormalized curve. Then

$$\frac{\gamma'(t)}{||\gamma'(t)||} = T(s(t)).$$

Differentiate this formula with respect to t using the fact that $||\gamma'(t)|| = \sqrt{\gamma'(t)\cdot\gamma'(t)}$ to obtain

$$\frac{||\gamma'||\,\gamma'' - \frac{\gamma'\cdot\gamma''}{||\gamma'||}\gamma'}{||\gamma'||^2} = \kappa N\,||\gamma'||$$

It follows that κ is the length of the vector

$$\frac{||\gamma'||\,\gamma'' - \frac{\gamma'\cdot\gamma''}{||\gamma'||}\gamma'}{||\gamma'||^3}$$

The square of this length can be obtained by dotting the vector with itself, which gives

$$\frac{||\gamma'||^2\,||\gamma''||^2 - (\gamma'\cdot\gamma'')^2}{||\gamma'||^6}$$

But the numerator of this fraction is $||\gamma' \times \gamma''||^2$, so we obtain the formula

$$\kappa = \frac{||\gamma' \times \gamma''||}{||\gamma'||^3}$$

Dividing out this κ from the previously displayed formula involving N gives

$$N = \frac{||\gamma'||}{||\gamma' \times \gamma''||} \left(\gamma'' - \frac{\gamma'}{||\gamma'||} \left(\frac{\gamma'}{||\gamma'||} \cdot \gamma'' \right) \right)$$

We get a simplier expression by crossing this expression with $T = \frac{\gamma'}{||\gamma'||}$ since $\gamma' \times \gamma' = 0$. But $N \times T = -T \times N = -B$. So

$$B(t) = \frac{\gamma' \times \gamma''}{||\gamma' \times \gamma''||}.$$

Finally, write $B(s(t))\,||\gamma' \times \gamma''|| = \gamma' \times \gamma''$. Differentiate both sides to obtain

$$\left(-\tau N \,||\gamma'||\right)\; ||\gamma' \times \gamma''|| + B\frac{(\gamma' \times \gamma''') \cdot (\gamma' \times \gamma'')}{||\gamma' \times \gamma''||} = \gamma' \times \gamma'''$$

Dot both sides with N to obtain

$$-\tau \,||\gamma'||\; ||\gamma' \times \gamma''|| = N \cdot (\gamma' \times \gamma''').$$

Apply the previous formula to N, noticing that $\gamma' \cdot (\gamma' \times \gamma''') = 0$, to obtain

$$-\tau \,||\gamma'||\; ||\gamma' \times \gamma''|| = \frac{||\gamma'||}{||\gamma' \times \gamma''||} \gamma'' \cdot (\gamma' \times \gamma'')$$

and so

$$\tau = -\frac{(\gamma' \times \gamma''') \cdot \gamma''}{||\gamma' \times \gamma''||^2} = \frac{(\gamma' \times \gamma'') \cdot \gamma'''}{||\gamma' \times \gamma''||^2}.$$

1.12 Higher Dimensions

Finally, I'd like to sketch the theory of curves in higher dimensions. Suppose $\gamma(s)$ is a curve parameterized by arclength in R^n. Consider the n-dimensional vectors $X_1 = \frac{d\gamma}{ds}, X_2 = \frac{d^2\gamma}{ds^2}, \ldots, X_n = \frac{d^n\gamma}{ds^n}$. Just as we assumed that $\kappa \neq 0$ in the three-dimensional case, we now assume that these vectors are always linearly independent.

Recall the Gram-Schmidt process, which replaces n-linearly independent vectors by n orthonormal vectors. Apply this process to $X_1, X_2, \ldots, X_n$, obtaining $Y_1, Y_2, \ldots, Y_n$. Recall that Y_1 is X_1 normalized to have length one, Y_2 is X_2 rotated in the plane generated by X_1 and X_2 until it is perpendicular to Y_1 and then normalized to have length one, etc. Since $\gamma(s)$ is parameterized by arclength, X_1 is already normalized and so $Y_1 = X_1$.

There is a geometric interpretation of the X_i or Y_i. Namely, the curve γ is momentarily trapped in the line spanned by Y_1, momentarily trapped in the plane spanned by Y_1 and Y_2, momentarily trapped in the 3-dimensional space spanned by Y_1, Y_2, Y_3, etc.

In the classical case, the curvature κ measures how rapidly the curve twists out of the line in which it finds itself momentarily trapped, and the torsion τ measures how rapidly the curve twists out of the osculating plane in which it finds itself momentarily trapped. Exactly the same thing happens in higher dimensions. We define a curvature κ_1 which measures how rapidly the curve twists out of the line in which it finds itself momentarily trapped. We define a higher curvature κ_2, which measures how rapidly the curve twists out of the two-dimensional plane in which it finds itself momentarily trapped. There are similar higher dimensional measurements $\kappa_3, \ldots, \kappa_{n-1}$.

It is easy to define κ_i and simultaneously prove an analogue of the Frenet-Serret formulas. Since the Y_i form a basis for R^n, we can write the derivative of each Y as a linear combination of the Y_i. Using the same notation as earlier, we have

$$\frac{dY_i}{ds} = \sum_j a_{ij} Y_j.$$

Because the Y_i are orthonormal, we conclude exactly as in section 1.6 that

$$a_{ij} = -a_{ji}.$$

In matrix notation, therefore, we have Frenet-Serret like formulas which look like

$$\frac{d}{ds}\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 0 & a_{12} & a_{13} & \ldots & a_{1n} \\ -a_{12} & 0 & a_{23} & \ldots & a_{2n} \\ -a_{13} & -a_{23} & 0 & \ldots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_{1n} & -a_{2n} & -a_{3n} & \ldots & 0 \end{pmatrix}\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix}$$

Now notice that the derivative of X_i is just X_{i+1} because $X_i = \frac{d^i\gamma}{ds^i}$. Therefore the derivative of X_1 is X_2 and thus a linear combination of Y_1 and Y_2. Since $Y_1 = X_1$, we conclude that the derivative of Y_1 is a linear combination of Y_1 and Y_2. But that means that the entire top row of the above matrix equation is zero except for a_{12}. We define κ_1 to be a_{12}.

In exactly the same way, we conclude that the derivative of Y_i is a linear combination of $Y_1, Y_2, \ldots, Y_{i+1}$. For example, the derivative of Y_2 is a linear combination of Y_1, Y_2, and Y_3.

But then every entry on the second row of the above matrix is zero except $-a_{12} = -\kappa_1$ and a_{23}. We define $\kappa_2 = a_{23}$. Etc.

In this way we kill two birds with one stone. We define the κ_i and simultaneously deduce that the Frenet-Serret formula becomes

$$\frac{d}{ds}\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 0 & \kappa_1 & 0 & 0 & \dots & 0 \\ -\kappa_1 & 0 & \kappa_2 & 0 & \dots & 0 \\ 0 & -\kappa_2 & 0 & \kappa_3 & \dots & 0 \\ 0 & 0 & -\kappa_3 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_n \end{pmatrix}$$

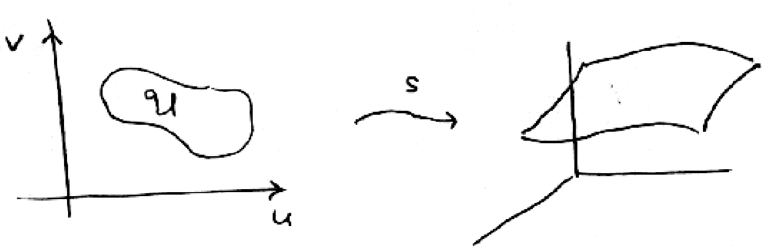
From here the theory proceeds exactly as before. We have $\frac{d\gamma}{ds} = Y_1$; this equation together with the Frenet-Serret formulas completely determines the curve up to initial conditions, and therefore the κ_i completely determine the curve.

Chapter 2

Surfaces

2.1 Parameterized Surfaces

Suppose we have a surface $\mathcal{S}$ inside R^3. Examples include the graph of $y^2 - x^2$, the surface of a sphere, the surface of a doughnut, and the Mobius band. A key feature of each of these surfaces is that in small pieces it looks like a piece of the plane. Thus whenever $p \in \mathcal{S}$, there is an open set $\mathcal{U}$ in the plane and a map s from $\mathcal{U}$ to the surface whose image is an open neighborhood of p in the surface.



There are two aspects to surface theory. In the *local theory,* we work with small pieces of a surface without worrying about global topology. In the *global theory,* we glue the information from the local theory together to get information about the entire surface. This chapter is about the local theory, so we will not worry when our symbolism does not describe every point on a surface.

Definition 7 *A* parameterized surface *is a* C^∞ *map* $s(u,v) : \mathcal{U} \subseteq R^2 \to R^3$ *defined on an open subset of the uv-plane such that*

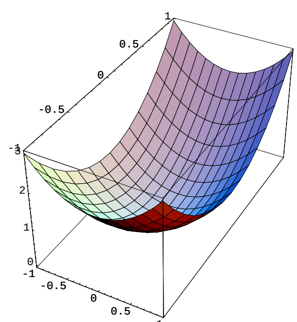
1. *s is one-to-one on $\mathcal{U}$*
2. *$\frac{\partial s}{\partial u}$ and $\frac{\partial s}{\partial v}$ are linearly independent at each point of $\mathcal{U}$, or equivalently $\frac{\partial s}{\partial u} \times \frac{\partial s}{\partial v} \neq 0$ at each point of $\mathcal{U}$.*

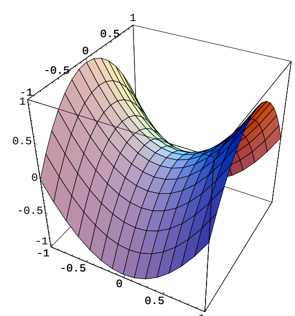
Refer again to the picture on the previous page. This picture is very, very important for the course. It describes the essential philosophy of differential geometry. Although we are interested in the surface, we will almost always work in the local coordinate system u, v. We'll do that so often that you may loose track of the fact that we are getting results which say something interesting about the surface.

Example 1: If $f(x,y)$ is an arbitrary C^∞ function, the graph of this function becomes a parameterized surface by writing

$$s(u,v) = (u, v, f(u,v)).$$

The local coordinates of a point can be found by projecting the point down to the plane. Notice that $\frac{\partial s}{\partial u} = \left(1, 0, \frac{\partial f}{\partial u}\right)$ and $\frac{\partial s}{\partial v} = \left(0, 1, \frac{\partial f}{\partial v}\right)$ point in different directions. Below are favorite examples generated by $x^2 + y^2$ and $y^2 - x^2$.

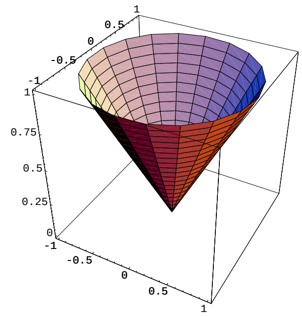




Example 2: Let $s(r,\theta) = (r\cos\theta, r\sin\theta, r)$. Clearly the resulting surface is $z = \sqrt{x^2+y^2}$ and so the cone below. Notice that this cone has a sharp singularity at the origin and yet the map s is C^∞. The partial derivatives of s are

$$\frac{\partial s}{\partial r} = (\cos\theta, \sin\theta, 1) \qquad\qquad \frac{\partial s}{\partial \theta} = (-r\sin\theta, r\cos\theta, 0)$$

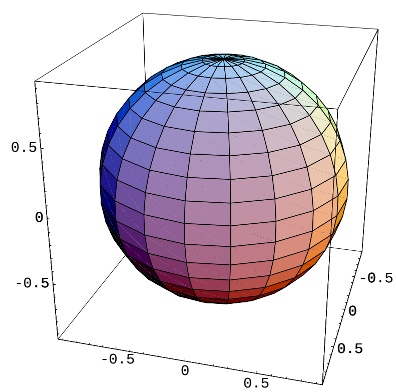
and at the origin these vectors are $(0,0,1)$ and $(0,0,0)$ and thus not linearly independent. This shows why we require that these two vectors be linearly independent.



Example 3: Standard spherical coordinates make the sphere into a parameterized surface.

$$s(\varphi, \theta) = (\sin\varphi\cos\theta, \sin\varphi\sin\theta, \cos\varphi)$$

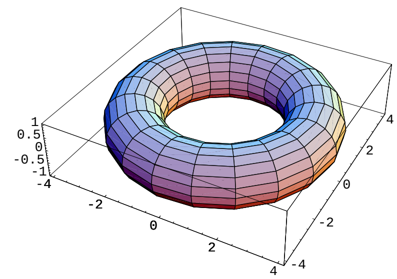
Recall that φ is the angle from the north pole down to the point in question, and θ is the angle which the projection of this point to the xy-plane makes with the x-axis. This map is not globally one-to-one on $0 \le \varphi \le \pi, 0 \le \theta \le 2\pi$ but is one-to-one on the inside of this rectangle. The inside maps to every point on the sphere except the north and south poles and the Greenwich meridian. Except at the poles, we have $\frac{\partial s}{\partial \varphi} \times \frac{\partial s}{\partial \theta} \neq 0$. Since we are only interested in local theory, we are happy to work with this coordinate map.



Example 4: Consider the doughnut shown on the next page. Let the radius of the circle through the center of the doughnut be R and let the radius of the smaller circle through a piece of doughnut be r. If p is a point on the doughnut, project p to the xy-plane and let θ be the angle made with the x-axis. Cut the doughnut alone the plane through the center and p and let the angle made from the center of the smaller circle to p be φ. It is easy to

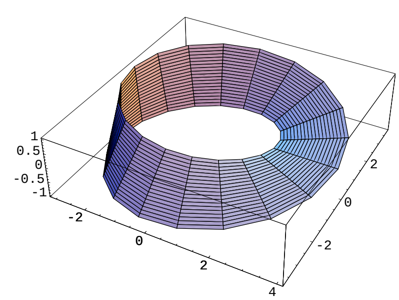
deduce the following parameterization:

$$s(\theta, \varphi) = ((R + r\cos\varphi)\cos\theta, (R + r\cos\varphi)\sin\theta, r\sin\varphi)$$



Example 5: Consider the Mobius band shown below. Let the center of this band trace out a circle $(R\cos\theta, R\sin\theta, 0)$. At each point of this circle, imagine a small arrow perpendicular to the circle and let this arrow rotate half as fast as the circle, so that when we go completely around the band, the arrow has twisted through 180 degrees. Clearly the arrow is $((R\cos\theta)\cos\theta/2, (R\sin\theta)\cos\theta/2, \sin\theta/2)$. We can get a band by going along the arrow a distance t, where $-1 \leq t \leq 1$. This gives the parameterization below.

$$s(\theta, t) = (R\cos\theta + tR\cos\theta\cos\theta/2, R\sin\theta + tR\sin\theta\cos\theta/2, t\sin\theta/2)$$



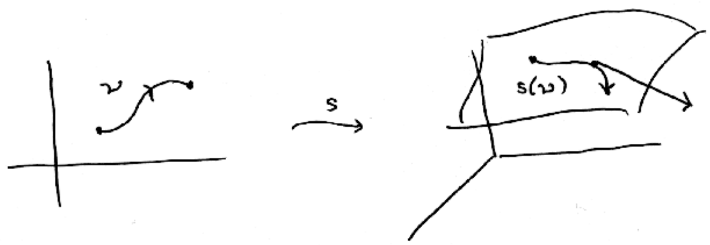
2.2 Tangent Vectors

Suppose we have a surface $\mathcal{S}$ and $p \in \mathcal{S}$. We want to describe the vectors in R^3 which are tangent to the surface at p. An easy way to obtain such vectors is to draw a curve

$\gamma(t) = (u(t), v(t))$ in local coordinates, move the curve to the surface by writing $s(u(t), v(t))$, and then differentiate. By the chain rule we get

$$\frac{\partial \vec{s}}{\partial u}\frac{du}{dt} + \frac{\partial \vec{s}}{\partial v}\frac{dv}{dt}.$$

Here we have written small arrows over s to indicate that $\frac{\partial \vec{s}}{\partial u}$ and $\frac{\partial \vec{s}}{\partial v}$ are three-dimensional vectors, but we will usually omit these arrows.



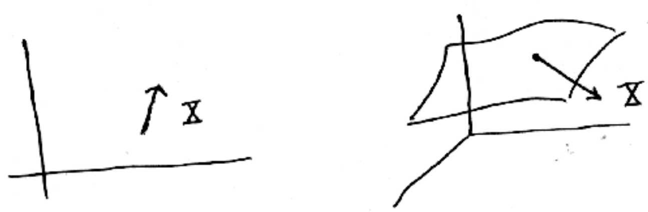
Notice that the expression just written is a linear combination of the linearly independent vectors $\frac{\partial \vec{s}}{\partial u}$ and $\frac{\partial \vec{s}}{\partial v}$. This motivates the following definition:

Definition 8 *Let $\mathcal{S}$ be a parameterized surface, $p \in \mathcal{S}$. A* tangent vector *at p is a vector in R^3 of the form*

$$X = X_1 \frac{\partial \vec{s}}{\partial u} + X_2 \frac{\partial \vec{s}}{\partial v}$$

where X_1 and X_2 are real numbers and the derivatives of s are evaluated at the point (u_0, v_0) corresponding to p. The corresponding vector in local coordinates *is the vector in R^2 given by*

$$X = (X_1, X_2).$$



Example: Consider the saddle $z = y^2 - x^2$ and the point $p = (1, 2, 3)$ on this saddle. The saddle can be parameterized by $s(u, v) = (u, v, v^2 - u^2)$ and the point corresponds to $u = 1, v = 2$. We have $\frac{\partial s}{\partial u} = (1, 0, -2u)$ and $\frac{\partial s}{\partial v} = (0, 1, 2v)$ and at the point p these vectors are $(1, 0, -2)$ and $(0, 1, 4)$. So a tangent vector is any linear combination of these vectors. Such a vector has the form

$$X = (X_1, X_2, -2X_1 + 4X_2)$$

and corresponds to the local coordinate vector

$$X = (X_1, X_2).$$

Remark: Recall that we are interested in surfaces, but usually work in local coordinates. Consequently, when we think of a tangent vector X, we visualize a vector in 3-space tangent to the surface. But when we write the vector symbolically, we usually just write $X = (X_1, X_2)$. At rare moments, we need to remember that the corresponding vector tangent to the surface is $X_1 \frac{\partial \vec{s}}{\partial u} + X_2 \frac{\partial \vec{s}}{\partial v}$.

Tangent vectors usually appear in this course in one of two ways. We may have a curve in the surface and want to compute the curve's tangent vector, which will be tangent to the surface. Or we may be given a tangent vector X and want to compute the directional derivative $X(f)$ of a function f on the surface in the direction X. Here are more details about the first of these ideas.

Definition 9 *By a* curve *on a parameterized surface $\mathcal{S}$, we mean a curve $\alpha(t)$ in R^3 which has the form $\alpha(t) = s(\gamma(t))$ for a C^∞ curve $\gamma(t)$ in the coordinate uv-plane. This coordinate curve is often written $\gamma(t) = (u(t), v(t))$ and the corresponding surface curve is written*

$$\alpha(t) = s(\gamma(t)) = (x(u(t), v(t)),\ y(u(t), v(t)),\ z(u(t), v(t)))\,.$$

Suppose we want the tangent vector to such a curve. We claim that we can perform the obvious calculation in either coordinate space and get the same answer. Indeed, in R^3 we will compute

$$\gamma'(t) = \left(\frac{du}{dt}, \frac{dv}{dt}\right)$$

and in R^3 we will compute

$$\alpha'(t) = \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}\right).$$

But $\alpha(t) = s(\gamma(t))$, so by the chain rule we have $\alpha'(t) = \frac{\partial s}{\partial u}\frac{du}{dt} + \frac{\partial s}{\partial v}\frac{dv}{dt}$. Thus the derivative in coordinate space is $\left(\frac{du}{dt}, \frac{dv}{dt}\right)$ and the derivative in R^3 is $\frac{du}{dt}\frac{\partial \vec{s}}{\partial u} + \frac{dv}{dt}\frac{\partial \vec{s}}{\partial v}$. These are different descriptions of the same tangent vector because according to our earlier rule the vector (X_1, X_2) in R^2 corresponds to the vector $X_1 \frac{\partial \vec{s}}{\partial u} + X_2 \frac{\partial \vec{s}}{\partial v}$ in R^3.

Example: Suppose we parameterize the surface $z = x^2 + y^2$ using polar coordinates. Then $s(r, \theta) = (r\cos\theta, r\sin\theta, r^2)$. Consider the curve $\gamma(t) = (1, t)$ in local coordinates.

Then $\alpha(t) = s(\gamma(t)) = (\cos t, \sin t, 1)$. At the special time $t = \pi/4$, $\gamma' = (0,1)$ and $\alpha' = (-1/\sqrt{2}, 1/\sqrt{2}, 0)$. These vectors are equivalent because $(0,1)$ corresponds to $0 \cdot \frac{\partial \vec{s}}{\partial r} + 1 \cdot \frac{\partial \vec{s}}{\partial \theta} = (-r\sin\theta, r\cos\theta, 0)|_{\pi/4} = (-1/\sqrt{2}, 1/\sqrt{2}, 0)$.

2.3 Directional Derivatives

In vector calculus, the directional derivative $D_X f$ of a function f in the direction X gives the slope of f in that direction. Usually X is a unit vector and it is proved that $D_X f = X \cdot \mathrm{grad} f = \sum_i X_i \frac{\partial f}{\partial x_i}$.

In differential geometry, these directional derivatives are very important, so we adopt an easier notation. We will always write $X(f)$ rather than $D_X f$. Moreover, we do not insist that X be a unit vector, defining $X(f) = \sum_i X_i \frac{\partial f}{\partial x_i}$ for any X. Of course this only gives a slope when X is a unit vector.

Following our usual philosophy, we search for a way to compute this derivative in local coordinates. If $f(x,y,z)$ is a function on R^3, then $f \circ s(u,v)$ is the corresponding function in u and v. We call this function $g(u,v)$. Happily, in local coordinates we obtain the nicest possible result:

Theorem 11 *Suppose $X = (X_1, X_2)$ is a two-dimensional vector in the uv-plane, corresponding to a three dimensional vector $X = X_1 \frac{\partial \vec{s}}{\partial u} + X_2 \frac{\partial \vec{s}}{\partial v}$ in R^3. Then $X(f)$, computed in R^3, is equal to the following expression computed in R^2 :*

$$X(g) = X_1 \frac{\partial g}{\partial u} + X_2 \frac{\partial g}{\partial v}$$

Proof: This is just the chain rule, since $X(f)$ equals

$$\sum_{i=1}^{2} X_i \frac{\partial g}{\partial u_i} = \sum_i X_i \frac{\partial (f \circ s)}{\partial u_i} = \sum_{ij} X_i \left(\frac{\partial f}{\partial x_j} \frac{\partial s_j}{\partial u_i} \right) = \sum_{ij} \left(X_i \frac{\partial s_j}{\partial u_i} \right) \frac{\partial f}{\partial x_j}$$

and the last expression is $X(g)$. QED.

Recall that algebraists like to write vectors using basis vectors. Instead of $X = (X_1, X_2)$, an algebraist would write $X = X_1 \vec{e}_1 + X_2 \vec{e}_2$. Physicists also like this notation and write $X = X_1 i + X_2 j$. Differential geometers also like basis vectors. But instead of writing $\vec{e}_1$ and $\vec{e}_2$, or i and j, they write $\frac{\partial}{\partial u}$ and $\frac{\partial}{\partial v}$ and thus write tangent vectors in the uv-plane as

$$X = X_1 \frac{\partial}{\partial u} + X_2 \frac{\partial}{\partial v}$$

Here $\frac{\partial}{\partial u}$ has no independent meaning; it is just a formal symbol. But obviously the notation has been chosen so

$$X(f) = \left(X_1 \frac{\partial}{\partial u} + X_2 \frac{\partial}{\partial v} \right) f = X_1 \frac{\partial f}{\partial u} + X_2 \frac{\partial f}{\partial v}$$

2.4 Vector Fields

Many people teaching calculus for the first time think it is very easy and straightforward and then get caught on tricky points. The first such point occurs very early in the term. Students are carefully taught that the derivative is the slope of the curve and thus a number. Then the lecturer must give an example. A typical first example is $f(x) = x^3$ and the derivative is $3x^2$. Wait. That's a function.

The point is that mathematicians usually differentiate functions at many points simultaneously, producing the function $\frac{df}{dx}$. Similarly in several variable calculate we differentiate at many places, producing the functions $\frac{\partial f}{\partial u}$ and $\frac{\partial f}{\partial v}$.

On a surface, there are no preferred directions. So the natural generalization of differentiation at a point is $X(f)$ for a fixed vector X. If we want to differentiate at many places at once, we must replace the vector X by a vector field.

Definition 10 *Let $\mathcal{S}$ be a surface. A* vector field *on $\mathcal{S}$ is an assignment to each point $p \in \mathcal{S}$ of a tangent vector $X(p)$ at p, such that these vectors vary in a C^∞ manner from point to point.*

Again, we usually work in local coordinates. Then a vector field has the form

$$X = X_1(u, v) \frac{\partial}{\partial u} + X_2(u, v) \frac{\partial}{\partial v}$$

where X_1 and X_2 are C^∞ functions of u and v. Notice that we use the same notation for vectors and vector fields. In practice this is not confusing. If X is a vector field and g is a function, then $X(g)$ is another function

$$X(g) = X_1(u, v) \frac{\partial f}{\partial u} + X_2(u, v) \frac{\partial f}{\partial v}$$

2.5 The Lie Bracket

The material in this section will not be used until the next chapter, so you may skip it and come back later if you wish.

In several variable calculus, it is very significant that mixed partial derivatives are equal, so $\frac{\partial^2 f}{\partial u \partial v} = \frac{\partial^2 f}{\partial v \partial u}$. On a surface there are no preferred directions, so we should work with arbitrary vector fields X and Y. But X and Y need not commute as operators. That is, $X(Y(g))$ need not equal $Y(X(g))$.

Non-commutative operations are one of mathematicians greatest discoveries. When operators do not commute, it is important to *measure* their noncommutativity. Rubic's cube is solved by operations of the form $aba^{-1}b^{-1}$, which measure the noncommutativity of simple twists a and b. In physics, measurement of noncommutativity leads to the Heisenberg uncertainty relations.

Also in this course, the deepest results will ultimately depend on measuring the noncommutativity of various operators. We have come to the first of these situations. If X and Y are vector fields, we define the Lie bracket (or Poisson bracket) of X and Y by the following formula.

Definition 11

$$[X, Y]g = X(Y(g)) - Y(X(g))$$

Notice that $[X, Y]$ makes no sense for vectors at a point; to define it, X and Y must be vector fields. The equality of mixed partial derivatives survives in the following remarkable theorem:

Theorem 12 *The operator $[X, Y]$ corresponds to a unique vector field.*

Proof: This theorem is a surprise because we expect that second derivatives would be involved. But they cancel out.

We have

$$X(Y(g)) = \left(X_1 \frac{\partial}{\partial u} + X_2 \frac{\partial}{\partial v}\right)\left(Y_1 \frac{\partial g}{\partial u} + Y_2 \frac{\partial g}{\partial v}\right)$$

Since the Y_i are functions, we must use the product rule. We obtain

$$\left(X_1 \frac{\partial Y_1}{\partial u} + X_2 \frac{\partial Y_1}{\partial v}\right)\frac{\partial g}{\partial u} + \left(X_1 \frac{\partial Y_2}{\partial u} + X_2 \frac{\partial Y_2}{\partial v}\right)\frac{\partial g}{\partial v} + \sum_{ij} X_i Y_j \frac{\partial^2 f}{\partial u_i \partial u_j}$$

Now subtract the same result with X and Y interchanged. Since mixed partial derivatives are equal, the last term cancels out and the final result is

$$\begin{aligned}[X, Y]g \quad = \quad & \left(X_1 \frac{\partial Y_1}{\partial u} + X_2 \frac{\partial Y_1}{\partial v} - Y_1 \frac{\partial X_1}{\partial u} - Y_2 \frac{\partial X_1}{\partial v}\right)\frac{\partial g}{\partial u} + \\ & \left(X_1 \frac{\partial Y_2}{\partial u} + X_2 \frac{\partial Y_2}{\partial v} - Y_1 \frac{\partial X_2}{\partial u} - Y_2 \frac{\partial X_2}{\partial v}\right)\frac{\partial g}{\partial u}\end{aligned}$$

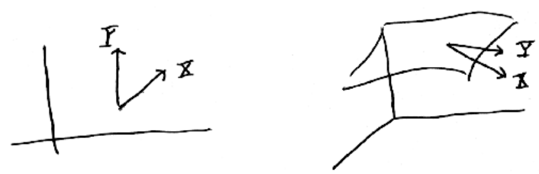
Thus $[X, Y]$ is obtained by applying another vector field, whose coefficients are given inside the large round brackets in the previous formula.

2.6 The Metric Tensor

The exciting ideas begin here. Suppose that X and Y are vectors in R^3. We can compute the dot product and length of these vectors. From now on, we will use the notation $\langle X, Y \rangle$ for dot product. Recall that $||X||^2 = \langle X, X \rangle$.

The basic philosophy of the course is to work in local coordinates. Suppose X and Y are vectors in the uv-plane. Then they correspond to tangent vectors to the surface X and Y, and we now define

Definition 12 *Suppose* $X = (X_1, X_2)$ *and* $Y = (Y_1, Y_2)$ *are vectors in* R^2. *Then* $\langle X, Y \rangle$ *denotes the dot product of the corresponding vectors in* R^3, *and* $||X|| = \sqrt{\langle X, X \rangle}$ *is the length of the corresponding vector in* R^3.



WARNING: These symbols *do not* mean ordinary dot product or length in two space.

Let us compute the formula for $\langle X, Y \rangle$. If $X = (X_1, X_2)$, then X corresponds to the three dimensional vector $X_1 \frac{\partial \vec{s}}{\partial u} + X_2 \frac{\partial \vec{s}}{\partial v}$. Similarly, Y corresponds to the vector $Y_1 \frac{\partial \vec{s}}{\partial u} + Y_2 \frac{\partial \vec{s}}{\partial v}$ and consequently, $\langle X, Y \rangle$ is the ordinary dot product of these vectors in R^3, which equals

$$(X_1 Y_1) \frac{\partial \vec{s}}{\partial u} \cdot \frac{\partial \vec{s}}{\partial u} + (X_1 Y_2 + X_2 Y_1) \frac{\partial \vec{s}}{\partial u} \cdot \frac{\partial \vec{s}}{\partial v} + (X_2 Y_2) \frac{\partial \vec{s}}{\partial v} \cdot \frac{\partial \vec{s}}{\partial v}$$

Definition 13

$$\begin{aligned} g_{11}(u, v) &= \frac{\partial \vec{s}}{\partial u} \cdot \frac{\partial \vec{s}}{\partial u} \\ g_{12}(u, v) &= \frac{\partial \vec{s}}{\partial u} \cdot \frac{\partial \vec{s}}{\partial v} \\ g_{22}(u, v) &= \frac{\partial \vec{s}}{\partial v} \cdot \frac{\partial \vec{s}}{\partial v} \end{aligned}$$

We also let $g_{21} = g_{12}$ and then we have

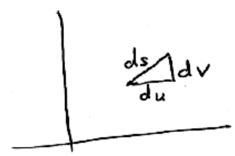
Theorem 13

$$\langle X, Y \rangle = \sum_{ij} g_{ij} X_i Y_j$$

$$||X|| = \sqrt{\sum_{ij} g_{ij} X_i X_j}$$

Because the g_{ij} are so important, we'll spend some time trying to get an intuitive feel for their significance. Suppose we are in the flat plane at a point $p = (u, v)$ and we make infinitesimal changes du and dv in u and v. Let ds be the distance we moved, that is, the hypotenuse. By Pythagoras,

$$ds^2 = du^2 + dv^2$$



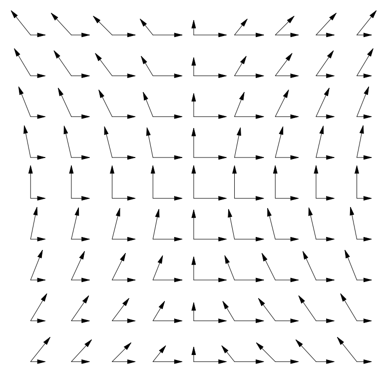
Now suppose that we are on a curved surface. We still draw coordinates u and v, but these coordinates no longer preserve lengths and angles. So the Pythagorian theorem fails. Gauss and Riemann's great idea was to replace the Pythagorian theorem with

$$ds^2 = \sum_{ij} g_{ij} du_i du_j$$

Here and in the future, we use u_1 and u_2 rather than u and v when we want to sum over various indices.

In the next section, we will use the g_{ij} to compute the lengths of curves on the surface in local coordinates. As we'll see, the infinitesimals above are replaced by calculations involving vectors.

Let's try a concrete example. Consider the saddle $z = y^2 - x^2$. Then $s(u, v) = (u, v, v^2 - u^2)$. Local coordinates are obtained by projecting points on the saddle straight down to the plane, but when arrows tangent to the saddle are projected down, their lengths change and angles between vectors change. We are going to use the mathematics above to find two vector fields X and Y in the plane which come from unit length perpendicular vector fields X and Y on the saddle. See the picture below.



Here are details. Since $\frac{\partial s}{\partial u} = (1, 0, -2u)$ and $\frac{\partial s}{\partial v} = (0, 1, 2v)$, we have $g_{11} = 1 + 4u^2, g_{12} = -4uv, g_{22} = 1 + 4v^2$.. Consider the vector field $X = (1, 0)$. While these vectors have length 1 in Euclidean space, their length on the surface is

$$\sqrt{\sum_{ij} g_{ij} X_i X_j} = \sqrt{g_{11}}$$

Consequently, the new vector field

$$\left(\frac{1}{\sqrt{g_{11}}}, 0 \right)$$

will represent a field of unit vectors on the saddle. These are the horizontal vectors in the previous picture. Notice that their Euclidean lengths vary from point to point.

We now attempt to find a second perpendicular vector field $Y = (a, b)$. The dot product of Y with $X = (1, 0)$ is $\sum g_{ij} X_i Y_j = g_{11} a + g_{12} b$. We want this to vanish, so we must choose $a = -\left(g_{12}/g_{11} \right) b$. We conclude that the second vector field should be a multiple of

$$\left(-\frac{g_{12}}{g_{11}}, 1 \right).$$

The length of this vector squared is

$$g_{11} \left(-\frac{g_{12}}{g_{11}} \right)^2 + 2g_{12} \left(-\frac{g_{12}}{g_{11}} \right) + g_{22} = \frac{g_{11} g_{22} - g_{12}^2}{g_{11}},$$

so we should choose the second vector field to be

$$\sqrt{\frac{g_{11}}{g_{11} g_{22} - g_{12}^2}} \left(-\frac{g_{12}}{g_{11}},\ 1 \right)$$

Remark: At the start of this course, we talked about a two-dimensional worker on the surface with an infinitesimal straightedge and compass, who could not see into the third dimension. Notice that such a worker could construct vector fields exactly as we have done. The worker would first draw coordinate lines on the surface. Experience would convince the worker that it is hopeless to try to draw perpendicular coordinates, or to be fussy about lengths, so the coordinates would look curved to us. Call the coordinate numbers u and v. The worker could use the ruler to draw unit vectors at each point, and the protractor and ruler to draw perpendicular unit vectors at each point.

A little thought shows that the worker could compute g_{ij}. Conversely, any calculation which the worker could perform with compass and ruler could be done by us if we knew g_{ij}.

So the g_{ij} contain exactly the information which could be discovered by a two-dimensional person who could not see into the third dimension.

There is one final remark to be made. We are assuming that there is a surface "out there." The g_{ij} describe ordinary Euclidean geometry on this surface. Our worker lives in coordinate space u, v and measures g_{ij} rather than computing them from the surface. If you wish to think this way, our worker lives in Plato's cave and only imagines the actual surface in the sunlight outside the cave.

But what if this surface didn't exist? Could we imagine that we measure g_{ij}, work in local coordinates, and do the rest of the calculations in this chapter without ever referring to the surface? Absolutely. That's exactly what we're going to do. And then might we later discover that there is no corresponding surface. Again, absolutely, this can happen. When it happens, we will have discovered a *new geometry* rather than just mundanely studying known surfaces.

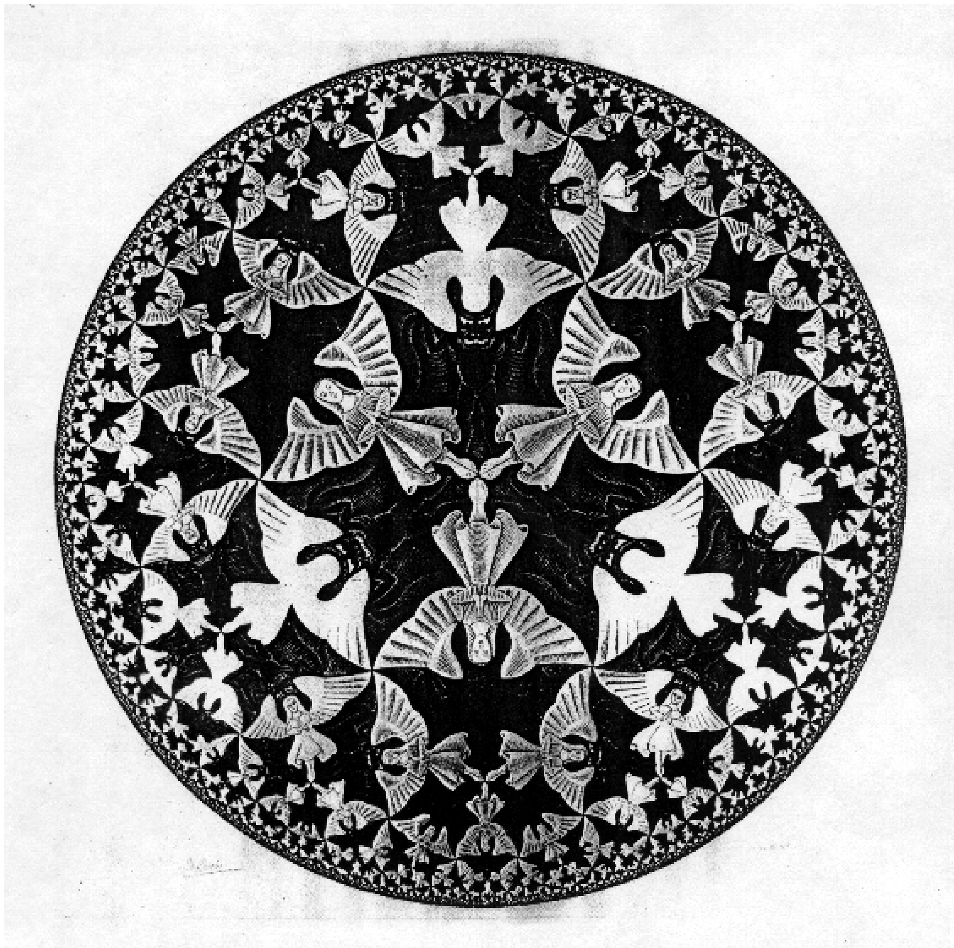
Here is an example. Define

$$ds^2 = \frac{4\ \left(dx^2 + dy^2\right)}{\left(1 - (x^2 + y^2)\right)^2}$$

on the unit disk. The four is included for historical reasons. You can immediately deduce the corresponding g_{ij}. When we are at the center of the disk, this is just the ordinary Pythagorean formula (except for the 2). But near the boundary, the denominator is almost zero and distances are *much later* than they appear.

It turns out that this example is exactly the non-Euclidean geometry discovered by Bolyai and Lobachevsky. It also turns out that no surface in R^3 can give these g_{ij} (this is a theorem of Hilbert). So ds^2 determines a new geometry.

However, many people know this geometry, because it is the geometry behind Escher's picture of angels and devils. In the picture, each angel has the same non-Euclidean size, although they shrink in the Euclidean world. See the next page.



2.7 Geodesics and Curve Length

After these extensive preparations, we are ready to discuss a truly great result. Let $\gamma(t) = (u(t), v(t))$ be a curve from a point p to a point q, written in local coordinates. We call this curve a *geodesic* if it is the shortest curve from p to q. Of course we are really measuring lengths on the surface, so geodesics will not be straight lines.

We are going to prove that each geodesic satisfies a differential equation. To make efficient use of summation notation, we will call the coordinates u_1 and u_2 rather than u and v. In the special case when the surface is a flat plane, the differential equation will turn out to be

$$\frac{d^2 u_k}{dt^2} = 0$$

and geodesics will be straight lines. This is no surprise. The key question is how to generalize the above simple equation; after much work, mathematicians in the nineteenth century discovered that the appropriate generalization is

$$\frac{d^2 u_k}{dt^2} + \sum_{ij} \Gamma^k_{ij} \frac{du_i}{dt} \frac{du^j}{dt} = 0$$

where Γ^k_{ij}, the so-called Christoffel symbols, are given by the formula

$$\Gamma^l_{ij} = \frac{1}{2} \sum_k \left(g^{-1}\right)_{lk} \left\{ \frac{\partial g_{jk}}{\partial u_i} + \frac{\partial g_{ik}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_k} \right\}$$

Notice in particular that the differential equation ultimately depends on g_{ij}, so geodesics can be determined by a two-dimensional worker.

Let us now begin the rigorous development of the theory. Suppose $\gamma(t) = (u(t), v(t))$ is a curve on the surface in local coordinates. We define the length of this curve from a to b to be

$$L^b_a(\gamma) = \int_a^b \left|\left|\gamma'(t)\right|\right| \, dt.$$

Recall that when we are in local coordinates, the expression $||\gamma'(t)||$ always means the special inner product using g_{ij}. Thus this definition when written out gives

$$L^b_a(\gamma) = \int_a^b \sqrt{\sum_{ij} g_{ij} \frac{du_i}{dt} \frac{du_j}{dt}} \, dt$$

We claim that this expression is equal to the actual Euclidean length of the curve $\alpha(t) = s(\gamma(t))$ on the surface. To see this, suppose that $X = X_1 \frac{\partial}{\partial u} + X_2 \frac{\partial}{\partial v}$ and $Y = Y_1 \frac{\partial}{\partial u} + Y_2 \frac{\partial}{\partial v}$ are vectors in local coordinates. By definition, the dot product $\langle X, Y \rangle$ equals the ordinary dot product of the corresponding vectors in three space given by $X = X_1 \frac{\partial s}{\partial u} + X_2 \frac{\partial s}{\partial v}$ and $Y = Y_1 \frac{\partial s}{\partial u} + Y_2 \frac{\partial s}{\partial v}$. Therefore, the local coordinate expression $||\gamma'(t)|| = \left|\left| \frac{du}{dt} \frac{\partial}{\partial u} + \frac{dv}{dt} \frac{\partial}{\partial v} \right|\right|$ is the ordinary Euclidean length of $\frac{du}{dt} \frac{ds}{du} + \frac{dv}{dt} \frac{ds}{dv}$, a vector which equals $\frac{d}{dt} s(u(t), v(t))$ by the chain rule. Since $s(u(t), v(t)) = \alpha(t)$, we have

$$\int_a^b \left|\left|\gamma'(t)\right|\right| \, dt = \int_a^b \left|\left|\alpha'(t)\right|\right| \, dt$$

as desired. Notice that this is a deceptive formula. The length on the left is computed in a fancy way using the g_{ij}, while the length on the right side is computed in three space using the Pythagorian theorem.

2.8 Energy

There is an underlying problem with geodesics. If $\gamma(t)$ is a curve and we reparameterize the curve by traveling along it at varying speed, the length remains the same and the new curve is still a geodesic. So geodesics can be written with infinitely many different parameterizations, and it is not clear which parameterization is picked out by the differential equation. We will soon see that the differential equation produces geodesics which travel with constant speed.

The solution to the problem of infinitely many parameterizations is provided by the physicists. Suppose a particle of mass m travels along our curve. The kinetic energy of this particle is

$$\frac{1}{2} \text{ mass} \times \text{velocity}^2$$

Thus the expression below, called the *total energy*, represents the sum of the energies of the particle during the moments of its journey from p to q.

$$\frac{1}{2} \text{ mass } \times \int_a^b \left|\left|\gamma'(t)\right|\right|^2 \, dt$$

Suppose that the particle is smart and wants to spend as little kinetic energy as possible getting from p to q. It turns out that the particle should do this by going along a geodesic at constant speed. If instead the particle travels along a longer curve, or travels along a geodesic but goes fast part of the time and slow part of the time, then it will use up more kinetic energy.

Since we are not physicists, constants do not matter to us.

Definition 14 *Let $\gamma(t)$ be a curve in local coordinates for $a \le t \le b$. The* energy *of the curve, $\mathcal{E}(\gamma)$, is defined to be*

$$\mathcal{E}(\gamma) = \int_a^b \left|\left|\gamma'(t)\right|\right|^2 \, dt$$

Theorem 14 *We have*

$$\left(L_a^b(\gamma)\right)^2 \le (b-a) \; \mathcal{E}(\gamma)$$

This inequality is an equality exactly when the speed of the curve is constant.

Proof: Recall that the dot product of two vectors in R^n equals the length of one times the length of the other times the cosine of the angle between them. So we have

$$\langle X, Y \rangle = ||X|| \; ||Y|| \; \cos\theta$$

Since $\cos\theta$ is between -1 and 1, the above formula implies the Schwarz inequality

$$\langle X, Y \rangle^2 \le ||X||^2 \; ||Y||^2$$

Moreover, this will be an equality exactly when $\cos\theta = \pm 1$ and so when X and Y point in the same direction and one is a scalar multiple of the other.

We now work by analogy. Let V be the vector space of all continuous functions on the interval $[a, b]$. Imagine that each $f \in V$ is a sort of infinite dimensional vector, with one coordinate $f(t)$ for each $t \in [a, b]$. Then it is reasonable to compute the dot product $\langle f, g\rangle$ by multiplying coordinates $f(t)g(t)$ and then summing up over all t. Since there are infinitely many t, we get

$$\langle f, g\rangle = \int_a^b f(t)g(t)\ dt.$$

The analogy leads to

Lemma 1 (Schwarz Inequality) *If $f(t)$ and $g(t)$ are continuous functions on $[a, b]$, then*

$$\left(\int_a^b f(t)g(t)\ dt\right)^2 \leq \int_a^b f^2(t)\ dt\ \int_a^b g^2(t)\ dt$$

with equality if and only if $g(t)$ is a constant multiple of $f(t)$.

You may or may not find the analogy convincing, but this result can easily be proved from scratch.

We apply the Schwarz inequality when $f(t) = 1$ and $g(t) = ||\gamma'(t)||$. It gives

$$\left(\int_a^b 1 \cdot \big|\big|\gamma'(t)\big|\big|\, dt\right)^2 \leq \int_a^b 1^2\ dt\ \int_a^b \big|\big|\gamma'(t)\big|\big|\ dt$$

and so $L(\gamma)^2 \leq (b - a)\ \mathcal{E}(\gamma)$. Moreover, this an equality exactly if $||\gamma'(t)||$ is a multiple of 1 so $\gamma(t)$ has constant speed. QED.

Remark: If we have a curve $\gamma(t)$ defined for $a \leq t \leq b$, we can reparameterize the curve so it has constant speed. The length will not change. By changing this constant speed, we can assume that the curve is defined on the interval $0 \leq t \leq 1$. From now on, we work only with such curves.

Theorem 15 *Consider curves in $\mathcal{U}$ defined for $0 \leq t \leq 1$. Such a curve minimizes energy if and only if it minimizes length and has constant speed.*

Proof: Suppose that $\gamma(t)$ has constant speed and the smallest possible length. Since the speed is constant, the previous theorem gives $L(\gamma)^2 = \mathcal{E}(\gamma)$ If $\tau(t)$ is another curve, then we have

$$\mathcal{E}(\gamma) = L(\gamma)^2 \leq L(\tau)^2 \leq \mathcal{E}(\tau)$$

and so γ has smallest energy.

Conversely, suppose γ has smallest energy. We have $L(\gamma)^2 \leq \mathcal{E}(\gamma)$ with equality exactly when γ has constant speed. Therefore if γ does not have constant speed, we could reparameterize to get a curve with the same length but constant speed and so smaller energy. If there were a shorter curve τ, then we could reparameterize τ to have constant speed, and then

$$\mathcal{E}(\gamma) = L(\gamma)^2 > L(\tau)^2 = \mathcal{E}(\tau)$$

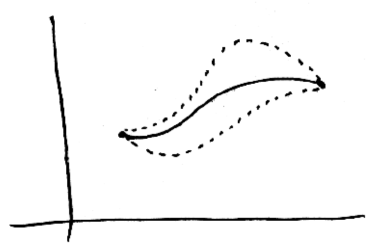
contradicting the assumption that γ has smallest energy. QED.

2.9 The Geodesic Equation

We now have enough background to deduce the differential equation of a geodesic. The calculation will be slightly messy, but after all we are deducing one of the great results in mathematics!

To avoid getting bogged down, we'll sketch the idea of the calculation first. Suppose we have a curve $\gamma(t)$ defined on $0 \leq t \leq 1$. Suppose the curve travels from p to q, so $\gamma(0) = p$ and $\gamma(1) = q$. Finally, suppose γ minimizes energy among such curves.

Imagine that we vary γ through a family of similar curves from p to q as in the picture below.

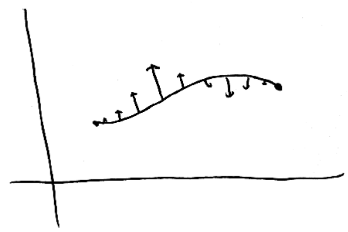


Call the family of curves $\gamma(t, s)$ where $-\epsilon < s < \epsilon$. For each fixed s we get a curve from p to q; when $s = 0$ we get our original curve γ.

Let $\mathcal{E}(s)$ be the energy of the s^{th} curve. Then $\mathcal{E}$ is a function of one variable which has a minimum at $s = 0$. By ordinary calculus, the derivative of $\mathcal{E}$ should be zero at $s = 0$. We are going to calculate this derivative by differentiating the formula for $\mathcal{E}$ with respect to s under the integral sign. It will turn out that the derivative at $s = 0$ has the form

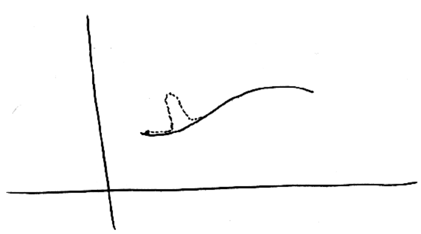
$$\int_0^1 \{ \text{ an expression in } t \text{ involving } g_{ij} \text{ and } u_i \ \} \ \frac{\partial s}{\partial u} \ dt$$

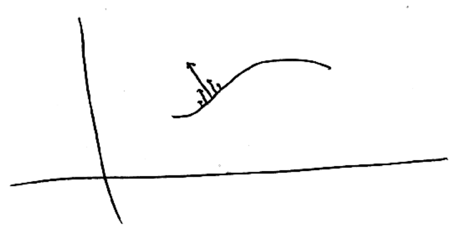
The expression $\frac{\partial s}{\partial u}(t,0)$ is the derivative of the variation, and thus a series of arrows along the original curve explaining how to begin the variation. See the picture below.



However, $\frac{d\mathcal{E}}{ds} = 0$ for all possible variations. Whenever we have a candidate $\frac{\partial s}{\partial u}(t,0)$ for a variation with the property that the arrows at zero at $t = 0$ and $t = 1$ so the varied curves still go from p to q, we can clearly find a variation $\gamma(t,s)$ with the appropriate derivative. So the integral must be zero for any function $\frac{\partial s}{\partial u}(t,0)$ which vanishes at the endpoints.

It follows that the expression "{ an expression in t involving g_{ij} and u_i}" inside the integral must be identically zero in t. Why? You might think that it could be nonzero but have positive and negative parts which cancel when integrated from 0 to 1. But since we can choose $\frac{\partial s}{\partial u}(t,0)$ arbitrarily, we could defeat this cancellation by choosing a function with a small bump, as illustrated below.





Finally, the expression "{ an expression in t involving g_{ij} and u_i}" will turn out to be our differential equation.

Fine. Here are the details. The energy of the variation $\gamma(t,s)$ is given by

$$\mathcal{E}(s) = \int_0^1 \left|\left| \frac{\partial \gamma}{\partial t}(t,s) \right|\right|^2 dt = \int_0^1 \sum_{ij} g_{ij}(u(t,s), v(t,s)) \frac{\partial u_i}{\partial t}(t,s) \frac{\partial u_j}{\partial t}(t,s)\; dt$$

We are assuming that the curve when $s = 0$ has smallest energy, so by calculus we have $\frac{d}{ds}\mathcal{E}(s) = 0$ when $s = 0$. Let us compute this derivative by integrating with respect to s

under the integral sign:

$$\frac{d\mathcal{E}}{ds} = \int_0^1 \sum_{ij} \frac{\partial}{\partial s}\left(g_{ij}(u(t,s),v(t,s))\right)\frac{\partial u_i}{\partial t}\frac{\partial u_j}{\partial t} + \sum_{ij} g_{ij}\frac{\partial^2 u_i}{\partial s \partial t}\frac{\partial u_j}{\partial t} + \sum_{ij} g_{ij}\frac{\partial u_i}{\partial t}\frac{\partial^2 u_j}{\partial s \partial t}$$

The first term can be expanded via the chain rule, so the entire expression becomes

$$\frac{d\mathcal{E}}{ds} = \int_0^1 \sum_{ijk} \frac{\partial g_{ij}}{\partial u_k}\frac{\partial u_k}{\partial s}\frac{\partial u_i}{\partial t}\frac{\partial u_j}{\partial t} + \sum_{ij} g_{ij}\frac{\partial^2 u_i}{\partial s \partial t}\frac{\partial u_j}{\partial t} + \sum_{ij} g_{ij}\frac{\partial u_i}{\partial t}\frac{\partial^2 u_j}{\partial s \partial t}$$

We now come to the decisive step. Notice that the first term in the previous integral is indeed a complicated expression multiplied by $\frac{\partial u_k}{\partial s}$ as promised. But the other two terms don't have this form. Instead they contain terms $\frac{\partial^2 u_i}{\partial s \partial t}$. We are going to integrate the last two terms by parts to convert to the required form. Recall that integration by parts is the formula $\int_a^b f(x)\frac{dg}{dx} = f(x)g(x)|_a^b - \int_a^b \frac{df}{dx}g(x)$. Here is the calculation for just these last two terms:

$$\int_0^1 \sum_{ij}\left\{[g_{ij}\frac{\partial u_j}{\partial t}\right\}\frac{\partial^2 u_i}{\partial t \partial s}\,dt + \int_0^1 \sum_{ij}\left\{g_{ij}\frac{\partial u_i}{\partial t}\right\}\frac{\partial^2 u_j}{\partial t \partial s}\,dt =$$

$$\text{(boundary terms)} - \int_0^1 \sum_{ij}\frac{\partial}{\partial t}\left\{[g_{ij}\frac{\partial u_j}{\partial t}\right\}\frac{\partial u_i}{\partial s}\,dt - \int_0^1 \sum_{ij}\frac{\partial}{\partial t}\left\{g_{ij}\frac{\partial u_i}{\partial t}\right\}\frac{\partial u_j}{\partial s}\,dt$$

Let us stop to investigate the boundary teams. The boundary term for the first integral is

$$\sum_{ij}\left\{[g_{ij}\frac{\partial u_j}{\partial t}\right\}\frac{\partial u_i}{\partial s}\Bigg|_0^1$$

However, $\gamma(t,s) = (u_1(t,s), u_2(t,s))$ is constantly p at $t = 0$ because all of our curves begin and end at p. So $\frac{\partial u_i}{\partial s} = 0$ when $t = 0$. Similarly this expression equals zero when $t = 1$. We conclude that all of our boundary terms vanish.

We'll now write the entire expression, factoring out $\frac{\partial u_k}{\partial s}$. To make this happen, we must rename a few indices and use the fact that $g_{ij} = g_{ji}$. We get

$$\frac{d\mathcal{E}}{ds} = \int_0^1 \sum_k \left\{\sum_{ij}\frac{\partial g_{ij}}{\partial u_k}\frac{\partial u_i}{\partial t}\frac{\partial u_j}{\partial t} - \sum_j \frac{\partial}{\partial t}\left[g_{jk}\frac{\partial u_j}{\partial t}\right] - \sum_i \frac{\partial}{\partial t}\left[g_{ik}\frac{\partial u_i}{\partial t}\right]\right\}\frac{\partial u_k}{\partial s}$$

At this point, we apply the second main idea of the proof. Since this expression must equal zero for all choices of $\frac{\partial u_k}{\partial s}$ which vanish at the endpoints, the integrand must be identically zero. So for each fixed k we have

$$\sum_{ij}\frac{\partial g_{ij}}{\partial u_k}\frac{\partial u_i}{\partial t}\frac{\partial u_j}{\partial t} - \sum_j \frac{\partial}{\partial t}\left[g_{jk}\frac{\partial u_j}{\partial t}\right] - \sum_i \frac{\partial}{\partial t}\left[g_{ik}\frac{\partial u_i}{\partial t}\right] = 0.$$

Expand out the derivative of the terms inside the square brackets, using the product rule and the chain rule. We get

$$\sum_{ij} \frac{\partial g_{ij}}{\partial u_k} \frac{\partial u_i}{\partial t} \frac{\partial u_j}{\partial t} - \sum_{ij} \frac{\partial g_{jk}}{\partial u_i} \frac{\partial u_i}{\partial t} \frac{\partial u_j}{\partial t} - \sum_j g_{jk} \frac{\partial^2 u_j}{\partial t^2} - \sum_{ij} \frac{\partial g_{ik}}{\partial u_j} \frac{\partial u_i}{\partial t} \frac{\partial u_j}{\partial t} - \sum_i g_{ik} \frac{\partial^2 u_i}{\partial t^2} = 0.$$

Notice that the two terms with second derivatives are really the same. They just use different indices. We write these terms together and use the fact that $g_{ik} = g_{ki}$ to get

$$\sum_{ij} \left\{ \frac{\partial g_{ij}}{\partial u_k} - \frac{\partial g_{jk}}{\partial u_i} - \frac{\partial g_{ik}}{\partial u_j} \right\} \frac{\partial u_i}{\partial t} \frac{\partial u_j}{\partial t} - 2 \sum_i g_{ki} \frac{\partial^2 u_i}{\partial t^2} = 0$$

Let g be the matrix

$$g = \begin{pmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{pmatrix}$$

The determinant of this matrix is positive. Indeed consider the tangent vectors in three-space defined by $X = \frac{\partial s}{\partial u}$ and $Y = \frac{\partial s}{\partial v}$. Recall also that $g_{11} = X \cdot X, g_{12} = X \cdot Y$, and $g_{22} = Y \cdot Y$. So the determinant $\det(g) = (X \cdot X)\ (Y \cdot Y) - (X \cdot Y)^2$ and by the Schwarz lemma this is nonnegative. Moreover, it is zero if and only if Y is a scalar multiple of X, but $\frac{\partial s}{\partial u}$ and $\frac{\partial s}{\partial v}$ point in different directions by definition of a surface.

Consequently, we can compute the inverse matrix g^{-1}. Multiply the previous equation by $-\frac{1}{2}(g^{-1})_{lk}$ and sum over k:

$$\sum_{ik} (g^{-1})_{lk} g_{ki} \frac{\partial^2 u_i}{\partial t^2} + \frac{1}{2} \sum_{ijk} (g^{-1})_{lk} \left\{ \frac{\partial g_{jk}}{\partial u_i} + \frac{\partial g_{ki}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_k} \right\} \frac{\partial u_i}{\partial t} \frac{\partial u_j}{\partial t} = 0.$$

Notice that $\sum_k (g^{-1})_{lk} g_{ki}$ is one if $l = i$ and zero otherwise. Consequently the above expression simplifies to

$$\frac{\partial^2 u_l}{\partial t^2} + \frac{1}{2} \sum_{ijk} (g^{-1})_{lk} \left\{ \frac{\partial g_{jk}}{\partial u_i} + \frac{\partial g_{ki}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_k} \right\} \frac{\partial u_i}{\partial t} \frac{\partial u_j}{\partial t} = 0.$$

Call the second expression Γ^l_{ij}. Also notice that we are interested only in the curve $\gamma(t, 0)$, so we can convert partial derivatives to total derivatives. We obtain

$$\frac{d^2 u_l}{dt^2} + \sum_{ij} \Gamma^l_{ij} \frac{du_i}{dt} \frac{du_j}{dt} = 0$$

We have proved the following theorem, where we change one index for convenience when we apply the result:

Theorem 16 *Define*

$$\Gamma_{ij}^{l} = \frac{1}{2}\sum_{k}(g^{-1})_{lk}\left\{\frac{\partial g_{jk}}{\partial u_i} + \frac{\partial g_{ik}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_k}\right\}$$

If $\gamma(t) = (u_1(t), u_2(t))$ *is a geodesic with constant speed, then*

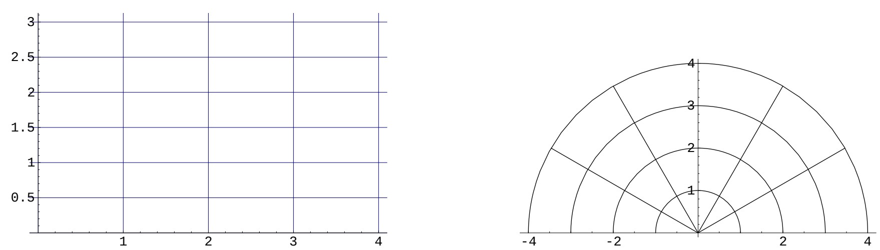
$$\frac{d^2u_k}{dt^2} + \sum_{ij}\Gamma_{ij}^{k}\frac{du_i}{dt}\frac{du_j}{dt} = 0.$$

2.10 An Example Where We Already Know the Answer

When I learn something new, I first like to try an example where I already know the answer to see if the new method really works.

Suppose we were to use polar coordinates in the plane rather than rectangular coordinates. Would the geodesic equation give the correct geodesics? Let's try.

We have a map $s(r,\theta) = (r\cos\theta, r\sin\theta, 0)$, illustrated below. Then $\frac{\partial s}{\partial r} = (\cos\theta, \sin\theta, 0)$ and $\frac{\partial s}{\partial \theta} = (-r\sin\theta, r\cos\theta, 0)$. So $g_{11} = \frac{\partial s}{\partial r}\cdot\frac{\partial s}{\partial r} = 1$, $g_{12} = \frac{\partial s}{\partial r}\cdot\frac{\partial s}{\partial \theta} = 0$ and $g_{22} = \frac{\partial s}{\partial \theta}\cdot\frac{\partial s}{\partial \theta} = r^2$.



Notice that

$$g = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \qquad \text{so} \qquad g^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{r^2} \end{pmatrix}$$

We must compute Γ_{ij}^{k}. Clearly $\Gamma_{ij}^{k} = \Gamma_{ji}^{k}$, so it suffices to compute the following expressions

$$\begin{aligned} \Gamma_{11}^{1} &= 0 \\ \Gamma_{11}^{2} &= 0 \\ \Gamma_{12}^{1} &= 0 \end{aligned}$$

$$\begin{aligned}\Gamma^2_{12} &= \frac{1}{r}\\ \Gamma^1_{22} &= -r\\ \Gamma^2_{22} &= 0\end{aligned}$$

Here's a hint about how these calculations were done. We have

$$\Gamma^l_{ij} = \frac{1}{2}\sum_k (g^{-1})_{lk}\left\{\frac{\partial g_{jk}}{\partial u_i} + \frac{\partial g_{ik}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_k}\right\}$$

In our particular case, g^{-1} has no offdiagonal elements, so the formula becomes

$$\Gamma^1_{ij} = \frac{1}{2}\left\{\frac{\partial g_{j1}}{\partial u_i} + \frac{\partial g_{i1}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_1}\right\} \qquad \Gamma^2_{ij} = \frac{1}{2r^2}\left\{\frac{\partial g_{j2}}{\partial u_i} + \frac{\partial g_{i2}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_2}\right\}$$

But the only g_{ij} term with nonzero derivative is g_{22} and the derivative is nonzero only if differentiated with respect to r, that is, to u_1. The result above immediately follows.

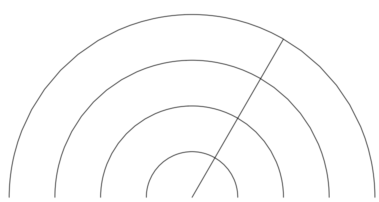
Next we write the geodesic equation. We have $\frac{d^2u_1}{dt^2} + \sum \Gamma^1_{ij}\frac{du_i}{dt}\frac{du_j}{dt} = 0$, which immediately translates to

$$\frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2 = 0$$

and we have $\frac{d^2u_2}{dt^2} + \sum \Gamma^2_{ij}\frac{du_i}{dt}\frac{du_j}{dt} = 0$, which immediately translates to

$$\frac{d^2\theta}{dt^2} + \frac{2}{r}\frac{dr}{dt}\frac{d\theta}{dt} = 0.$$

Suppose for a moment that a geodesic has constant θ. Then the two equations reduce to $\frac{d^2r}{dt^2} = 0$ and so $r = at + b$. This gives a radial line through the origin, clearly one kind of geodesic.



Otherwise θ varies and our geodesic can be given by an equation for r in terms of θ. We have

$$\begin{aligned}\frac{dr}{dt} &= \frac{dr}{d\theta}\frac{d\theta}{dt}\\ \frac{d^2r}{dt^2} &= \frac{d^2r}{d\theta^2}\left(\frac{d\theta}{dt}\right)^2 + \frac{dr}{d\theta}\frac{d^2\theta}{dt^2}\end{aligned}$$

and when these equations are substituted into the two earlier geodesic equations, we obtain

$$\begin{aligned} \frac{d^2r}{d\theta^2}\left(\frac{d\theta}{dt}\right)^2 + \frac{dr}{d\theta}\frac{d^2\theta}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2 &= 0 \\ \frac{d^2\theta}{dt^2} + \frac{2}{r}\frac{dr}{d\theta}\left(\frac{d\theta}{dt}\right)^2 &= 0 \end{aligned}$$

The second equation can be solved for $\frac{d^2\theta}{dt^2}$ and this result can be inserted into the first equation, yielding a single equation

$$\frac{d^2r}{d\theta^2}\left(\frac{d\theta}{dt}\right)^2 - \frac{2}{r}\left(\frac{dr}{d\theta}\right)^2\left(\frac{d\theta}{dt}\right)^2 - r\left(\frac{d\theta}{dt}\right)^2 = 0$$

and we can factor out a common term to give

$$\frac{d^2r}{d\theta^2} - \frac{2}{r}\left(\frac{dr}{d\theta}\right)^2 - r = 0$$

Incidentally, the reason our original two equations have become only a single equation is that we have lost track of how the curve is traced in time, and only kept information about the shape $r(\theta)$.

This final differential equation can be solved by an ingenious trick which also comes up in the theory of Newtonian orbits for the two body problem. Let $r(\theta) = \frac{1}{u(\theta)}$. Then $\frac{dr}{d\theta} = \frac{-1}{u^2}\frac{du}{d\theta}$ and $\frac{d^2r}{d\theta^2} = \frac{2}{u^3}\left(\frac{du}{d\theta}\right)^2 - \frac{1}{u^2}\frac{d^2u}{d\theta^2}$. So the previous equation becomes

$$\frac{2}{u^3}\left(\frac{du}{d\theta}\right)^2 - \frac{1}{u^2}\frac{d^2u}{d\theta^2} - 2u\left(\frac{-1}{u^2}\frac{du}{d\theta}\right)^2 - \frac{1}{u} = 0$$

and after miraculous cancellation the equation becomes

$$\frac{d^2u}{d\theta^2} + u = 0.$$

The solutions of this equation are $u(\theta) = A\sin(\theta + \delta)$ and so

$$r(\theta) = \frac{1}{A\sin(\theta + \delta)}$$

Luckily, this is the polar form of a straight line which misses the origin. Consider the pictures on the next page. The picture on the left shows that a horizontal line of distance $\frac{1}{A}$ from the origin has polar equation $r(\theta) = \frac{1}{A\sin\theta}$. The picture on the right shows that every straight line which misses the origin can be obtained by rotating the picture on the left by $-\delta$. When a polar curve $r(\theta)$ is rotated by $-\delta$, the new equation has the form $r(\theta + \delta)$.

Putting these results together, we find that the solutions of the geodesic equation in polar coordinates are exactly the polar forms of straight lines in the plane. Whew.

2.11 Geodesics on a Sphere

We end this chapter with a description of geodesics on several important surfaces. The geodesic equation is a nonlinear differential equation, and it can seldom be explicitly solved. Moreover, solutions of the geodesic equation have constant speed, and we have discovered that renormalizing a curve so it has constant speed is rarely possible explicitly. The equation can be solved numerically, however. It is not difficult to teach Mathematica how to do so; Mathematica has built-in routines to solve differential equations.

There are also tricks which can be useful!

Trick 1: The geodesic equation is a second order equation. Consequently, solutions $\gamma(t)$ are uniquely by the boundary conditions $\gamma(0)$ and $\gamma'(0)$. If a geodesic is already known with these boundary conditions, it must be the one solving the geodesic equations.

Trick 2: An isometry of a surface is a C^∞ map $M : \mathcal{S} \to \mathcal{S}$ which is one-to-one and onto with C^∞ inverse and preserves lengths. This final condition means that $M \circ \alpha(t)$ and $\alpha(t)$ always have the same length for any curve α. Such isometries automatically preserve geodesics (why?).

These tricks can be used to dramatically simplify the determination of geodesics on a sphere. One approach is to use spherical coordinates, write down the geodesic equation as in the previous section, and then solve in the special case that φ is a constant. The equations simplify dramatically and we discover that the solutions are circles which traverse the equator at constant speed. We call these solutions *great circles along the equator*.

Now suppose that we have a more general geodesic $\gamma(t)$ with boundary conditions $\gamma(0) = p$ and $\gamma'(0) = X$. Rotate the sphere so the equator rotates to a great circle through p in the direction X. By Trick 2, this new great circle must be a geodesic. By Trick 1, it must be the geodesic $\gamma(t)$. This proves

Theorem 17 *The geodesics on a sphere are exactly the great circles.*

Remark: This result is known to the general public. You'll see it dramatically confirmed if you fly to Europe and discover yourself over Greenland in the middle of the journey.

Remark: However, even the special calculation described above can be eliminated if we notice that the sphere also has isometries which reflect across the equator: $(x, y, z) \to (x, y, -z)$. Call this isometry M. Let $\gamma(t)$ be a geodesic which begins at the equator, so $\gamma(0) = p$ is on the equator, and initially moves tangent to the equator, so $\gamma'(0) = X$ is tangent to the equator. Then $M \circ \gamma$ is again a geodesic. Clearly this geodesic starts at p with initial velocity X. By Trick 1, it equals γ. But $M \circ \gamma(t)$ can equal $\gamma(t)$ only if $\gamma(t)$ has zero z-component. So γ is a great circle along the equator. QED.

Remark: It is very important to notice that the curve which follows the equator three fourths of the way around the sphere is a geodesic. It is certainly not the shortest curve connecting its endpoints, because it is shorter to go one fourth of the equator the other way. We proved that shortest curves are geodesics, but did not prove the converse.

It turns out that geodesics locally minimize distance. Given t, there is an $\epsilon > 0$ such that γ is the shortest curve to all $\gamma(u)$ satisfying $|t - u| < \epsilon$.

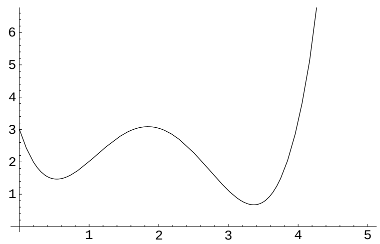
2.12 Geodesics on a Surface of Revolution

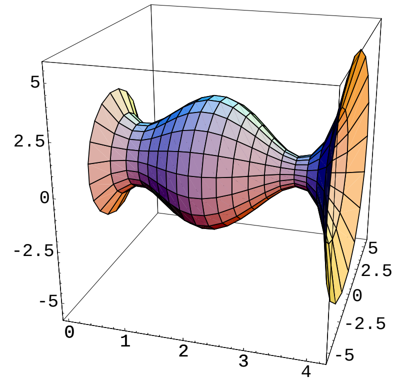
Consider the surface formed by rotating the graph of a function $y = f(x)$ about the x-axis. The resulting surface can be parameterized by

$$s(x, \theta) = (x, f(x) \cos\theta, f(x) \sin\theta)$$

Although we will not be able to solve the geodesic equation completely, we get a remarkably complete and beautiful description of the geodesics on a surface of revolution by using a combination of symbolic calculation and geometric interpretation.

When we draw pictures, we will draw x horizontally, thinking of y as going up and z as coming out at us. See the pictures on the next page.





We now deduce the geodesic equation. Notice that

$$\begin{aligned}
\frac{\partial s}{\partial x} &= (1, f'(x)\cos\theta, f'(x)\sin\theta) \\
\frac{\partial s}{\partial \theta} &= (0, -f(x)\sin\theta, f(x)\cos\theta) \\
g_{11} &= 1 + \left(f'(x)\right)^2 \\
g_{12} &= 0 \\
g_{22} &= (f(x))^2 \\
\Gamma^1_{11} &= \frac{f'f''}{1+(f')^2} \\
\Gamma^1_{12} &= 0 \\
\Gamma^1_{22} &= -\frac{ff'}{1+(f')^2} \\
\Gamma^2_{11} &= 0
\end{aligned}$$

$$\Gamma^2_{12} = \frac{f'}{f}$$

$$\Gamma^2_{22} = 0$$

We conclude that geodesics satisfy the following equations

$$\frac{d^2x}{dt^2} + \frac{f'f''}{1+(f')^2}\left(\frac{dx}{dt}\right)^2 - \frac{ff'}{1+(f')^2}\left(\frac{d\theta}{dt}\right)^2 = 0$$

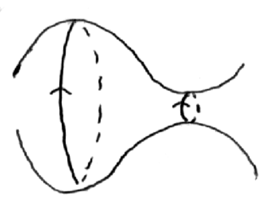
$$\frac{d^2\theta}{dt^2} + \frac{2f'}{f}\frac{dx}{dt}\frac{d\theta}{dt} = 0$$

Example 1: Suppose x is constant, so only θ varies. We will call such a curve a *meridian.* The equations reduce to

$$-\frac{ff'}{1+(f')^2}\left(\frac{d\theta}{dt}\right)^2 = 0$$

$$\frac{d^2\theta}{dt^2} = 0$$

The solution of the second equation is $\theta(t) = ct+d$, so the geodesic goes around the surface of revolution at constant angular velocity. If the geodesic is not constant, the first equation implies that $f'(x) = 0$. So meridians are geodesics if and only if x is a local min, local max, or other critical point of f.



It may seem strange that meridians for a local *maximum* of f are geodesics. But recall that geodesics minimize length only locally. The point is that we should move around the rim to get to nearby points as fast as possible.

Example 2: Suppose that θ is constant. I like to call such curves *latitudes.* The geodesic equations reduce to

$$\frac{d^2x}{dt^2} + \frac{f'f''}{1+(f')^2}\left(\frac{dx}{dt}\right)^2 = 0$$

At first this seems difficult to solve. However, geodesics have constant speed. Our particular geodesic is $(x(t), f(x(t)), 0)$, possibly rotated around the surface by a fixed θ. So its tangent vector is $\left(\frac{dx}{dt}, f'(x)\frac{dx}{dt}, 0\right)$ and the square of the length of this vector is

$$\left(1 + (f')^2\right)\left(\frac{dx}{dt}\right)^2$$

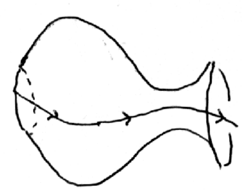
This number will be a constant just in case its derivative is zero, so

$$\left(2f'f''\frac{dx}{dt}\right)\left(\frac{dx}{dt}\right)^2 + \left(1 + (f')^2\right)2\frac{dx}{dt}\frac{d^2x}{dt^2} = 0$$

or

$$2\left(1 + (f')^2\right)\frac{dx}{dt}\left\{\frac{d^2x}{dt^2} + \frac{f'f''}{1 + (f')^2}\left(\frac{dx}{dt}\right)^2\right\} = 0$$

Notice that the expression inside the curley brackets is exactly the geodesic equation. So the geodesic equation merely states that the latitude must be traversed at constant speed.



We have proved

Theorem 18 *Meridians are geodesics exactly at critical points of $f(x)$. All latitudes are geodesics.*

We must now find the remaining geodesics, which wind around the surface. We begin by solving the second of the two geodesic equations:

$$\frac{d^2\theta}{dt^2} + \frac{2f'}{f}\frac{dx}{dt}\frac{d\theta}{dt} = 0$$

Multiply this equation by f^2 and replace $\frac{df}{dx}\frac{dx}{dt}$ by $\frac{d}{dt}f(x(t))$. The equation becomes

$$f^2\,\frac{d^2\theta}{dt^2} + 2f\,\frac{d}{dt}f(x(t)\,\frac{d\theta}{dt} = 0.$$

Equivalently, then

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

So the second equation just asserts that $f^2\frac{d\theta}{dt}$ is constant! Call this constant c_1. Then

$$\frac{d\theta}{dt} = \frac{c_1}{f(x)^2} \qquad \text{or} \qquad f(x)\frac{d\theta}{dt} = \frac{c_1}{f(x)}$$

Remark: Notice that $\frac{d\theta}{dt}$ is always positive or always negative or always zero. So once a curve starts winding around the surface of revolution clockwise, it always winds in that direction. It cannot stop winding, or start winding in the other direction.

The number $f(x)$ is the radius of the surface of revolution at x. If we watch our curve over a small interval of time dt, it moves along this radius a distance $f(x)d\theta$. It also moves perpendicularly to the radius along the curve $y = f(x)$ (or rather, along this curve rotated by θ). In the time dt, x will change by dx. But the curve $y = f(x)$ has slope f', so the distance our geodesic travels perpendicular to the radius is not dx but rather $\sqrt{1+(f')^2}\ dx$.

Recall that our curve has constant speed. If it travels along the radius by $f(x)\ d\theta$ and perpendicularly along the surface by $\sqrt{1+(f')^2}\ dx$, then the total distance it moves in time dt is

$$\sqrt{(f(x)\ d\theta)^2 + \left(\sqrt{1+(f')^2}\ dx\right)^2}$$

It is convenient to rewrite this

$$\sqrt{\left(f\frac{d\theta}{dt}\right)^2 + (1+(f')^2)\left(\frac{dx}{dt}\right)^2}\ dt$$

Since the curve has constant speed, the expression inside the square root must be a constant independent of t. Call this constant c_2^2. Our conclusions up to this point can be summarized in two equations:

$$f(x)\frac{d\theta}{dt} = \frac{c_1}{f(x)}$$

$$\left(f\frac{d\theta}{dt}\right)^2 + (1 + (f')^2)\left(\frac{dx}{dt}\right)^2 = c_2^2.$$

Remark: We now claim that this second equation is equivalent to the remaining unexamined geodesic equation. To see this, multiply that unexamined geodesic equation by $2\left(1 + (f')^2\right)\frac{dx}{dt}$ to get

$$2\left(1 + (f')^2\right)\frac{dx}{dt}\left\{\frac{d^2x}{dt^2} + \frac{f'f''}{1 + (f')^2}\left(\frac{dx}{dt}\right)^2\right\} - 2ff'\frac{dx}{dt}\left(\frac{d\theta}{dt}\right)^2 = 0.$$

We know from example 2 that the first of these two terms is the derivative with respect to t of $\left(1 + (f')^2\right)\left(\frac{dx}{dt}\right)^2$. Replace $\frac{d\theta}{dt}$ by $\frac{c_1}{f^2}$ in the second term to obtain $-2f\frac{df}{dt}\left(\frac{c_1}{f^2}\right)^2 = \frac{d}{dt}\left(\frac{c_1}{f}\right)^2$. So the unexamined geodesic equation states that the derivative of

$$\left(\frac{c_1}{f}\right)^2 + \left(1 + (f')^2\right)\left(\frac{dx}{dt}\right)^2$$

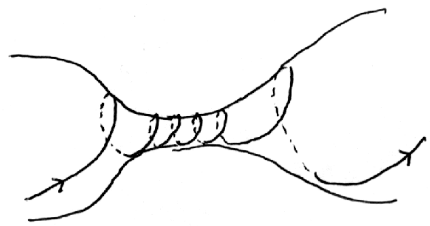
is zero and thus this expression is constant.

Remark: From now on, we try to understand the significance of the two equations

$$f(x)\frac{d\theta}{dt} = \frac{c_1}{f(x)}$$

$$\left(f\frac{d\theta}{dt}\right)^2 + (1 + (f')^2)\left(\frac{dx}{dt}\right)^2 = c_2^2.$$

According to the second equation, the motion of the geodesic is split into a component in the direction of the meridian circle through $\gamma(t)$ and a component in the direction of the latitude curve through $\gamma(t)$. As the radius decreases, the meridian component increases and the latitude component decreases, so the geodesic becomes *vertical*, winding tightly around the surface. As the radius increases, the meridian component decreases and the latitude component increases, so the geodesic becomes *horizontal*, winding very slowly around the surface.



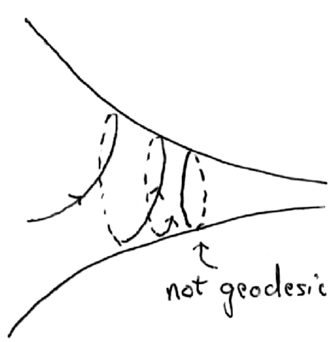
In addition, there is a minimal possible radius for the geodesic. Indeed,

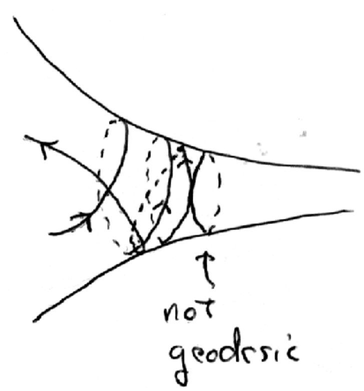
$$\left(\frac{c_1}{f}\right)^2 = \left(f\frac{d\theta}{dt}\right)^2 \leq c_2^2$$

and so

$$\frac{c_1}{c_2} \leq f(x).$$

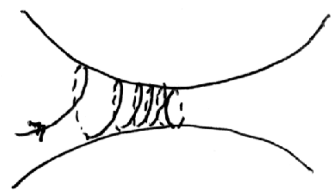
What happens when the geodesic approaches a spot on the surface of revolution where the radius is too small? The answer depends on whether this critical radius corresponds to a geodesic or not. In the picture below, the geodesic approaches a meridian at x where $f'(x) \neq 0$. This meridian is not a geodesic. As the geodesic approaches the meridian, it becomes more and more vertical. Finally the geodesic touches the meridian vertically and then bounces back toward the left.



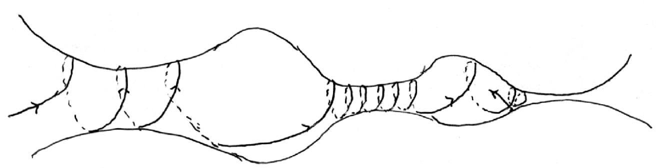


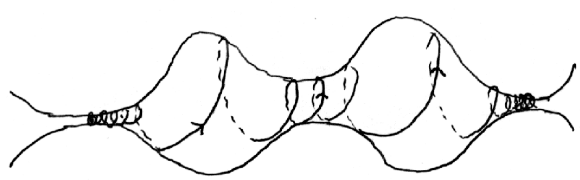
On the other hand, suppose the meridian of critical radius is a geodesic. If our geodesic approaching this critical radius were to touch the meridian, it would be vertical there.

By uniqueness of boundary conditions, our geodesic would have to equal the meridian. What happens instead is that our geodesic winds more and more vertically, approaching the meridian infinitely closely without ever touching or bouncing back.



Let us suppose that our surface of revolution extends infinitely far along the x-axis. Our analysis gives a complete picture of geodesics on the surface. These geodesics travels almost horizontally when the radius is large, and then pinch together as the radius decreases. If the radius does not decrease too much, the geodesic escapes through the narrow spot and continues on. If it finds a spot which is too narrow, it bounces back and retreats toward the left. And if it finds a spot which is just right, it approaches the spot infinitely closely. Some geodesics bounce back and forth forever. Others continue on, almost managing to squeeze through narrow spots.





2.13 Geodesics on a Poincare Disk

Finally, we will sketch the theory of geodesics on the Poincare disk described at the end of section 2.6. Since this geometry is exactly Lobachevsky's non-Euclidean geometry, it plays an important role in mathematics. We will not give complete details, but an interested reader can easily complete the theory from our sketch.

Recall that

$$ds^2 = \frac{4\,\left(dx^2+dy^2\right)}{\left(1-(x^2+y^2)\right)^2}$$

Notice first that reflection across a line through the origin preserves distance. Consequently the tricks deduced earlier allow us to conclude that geodesics through the origin remain on a radial line.

By rotational symmetry, it suffices to understand the curve $\gamma(t) = (x(t), 0)$. The same analysis will then apply to $r(t)$ for any radial curve.

Since geodesics travel at constant speed, the length of the derivative of our curve must remain constant. The resulting derivative is $\gamma'(t) = (x'(t), 0)$, and its Poincare length is

$$\frac{2}{1-x^2}\frac{dx}{dt}.$$

This must be a constant c, so $\frac{2\ dx}{1-x^2} = c\ dt$ and we discover by integration that $= ln\left(\frac{1+x}{1-x}\right) = ct + d$. By starting time at a different moment we can eliminate d. Exponentiating both sides gives $\left(\frac{1+x}{1-x}\right) = e^{ct}$ and so

$$x(t) = \frac{e^{ct}-1}{e^{ct}+1} = \frac{e^{ct/2}-e^{-ct/2}}{e^{ct/2}+e^{-ct/2}} = \tanh(ct/2).$$

Notice that as $t \to \infty$, $x(t) \to 1$.

Theorem 19 *Geodesics through the origin of the Poincare disk are radial lines which move more and more slowly as they approach the boundary and never reach the boundary.*

To find the remaining geodesics, we will show that there are a large number of isometries of the Poincare disk, so that we can get any geodesic by applying an isometry to a radial line.

For a moment, forget the Poincare disk and consider ordinary Euclidean geometry on the plane. We can think of this plane as the set of all complex numbers. Consider the function

$$f(z) = \frac{az+b}{cz+d}$$

which maps complex numbers to complex numbers; here a, b, c, and d are fixed complex numbers. Of course this map is not defined when $cz + d = 0$.

Lemma 2 *This map preserves angles in the sense that if $\alpha(t)$ and $\beta(t)$ are two curves in the plane which meet at an angle α, then $f(\alpha(t))$ and $f(\beta(t))$ meet at the same angle.*

Proof: This lemma can be proved by a brute force and also follows from general principles of complex variable theory.

Lemma 3 *The map f preserves straight lines and circles in the sense that the image of a straight line under f is either a straight line or a circle, and the image of a circle is either a straight line or a circle.*

Proof: Again this can be proved by brute force.

We now claim that certain of these maps send the unit circle centered at the origin back to itself, not necessarily fixing the circle pointwise. A brief calculation reveals that

Theorem 20 *Suppose z_0 is a fixed complex number satisfying $|z_0| < 1$, and θ is an arbitrary real number. The map $f(z)$ below sends the unit circle back to itself. Consequently this map sends the interior of the unit circle to itself. Indeed, this map is a one-to-one and onto map from the Poincare disk to itself, and its inverse has the same form for some other choice of z_0 and θ.*

$$f(z) = e^{i\theta} \frac{z - z_0}{1 - z\ \bar{z}_0}.$$

Finally, we come to the most important point, which can again be proved by brute force:

Theorem 21 *The map*

$$f(z) = e^{i\theta} \frac{z - z_0}{1 - z\ \bar{z}_0}.$$

from the Poincare disk to itself preserve Poincare length and so is an isometry of the disk.

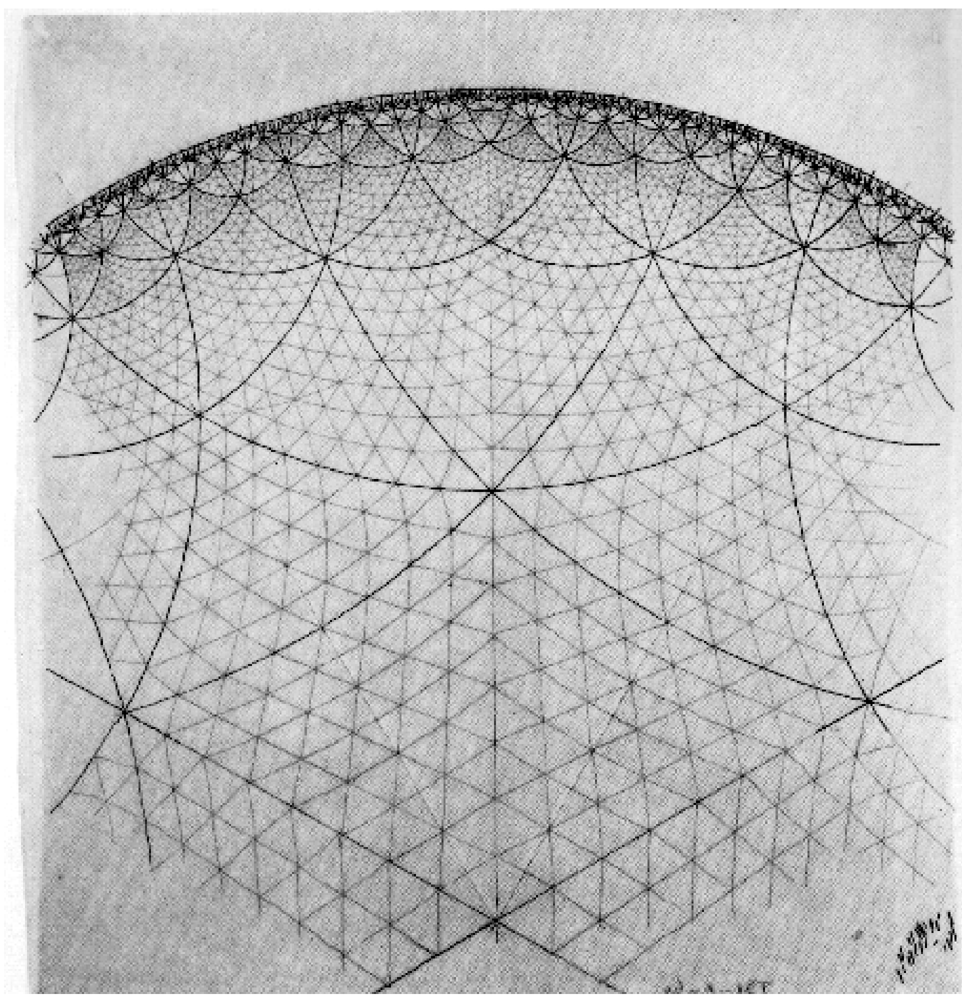
Turn back to Escher's picture of the Poincare disk in section 2.6. These isometries are clearly visible, mapping one angel to another and one devil to another.

Remark: We now have enough information to determine the geodesics of the Poincare disk completely. On the next page, these geodesics can be seen in Escher's working notes used to construct the angel and devil picture.

Theorem 22 *The geodesics on the Poincare disk are exactly radial lines through the origin or circles which meet the boundary at ninety degrees.*

Proof: Suppose a geodesic $\gamma(t)$ goes through the point z_0. Apply the isometry f given above. This isometry maps γ to a geodesic through the origin, and consequently to a radial line. This radial line meets the boundary of the disk at ninety degrees. Now apply the inverse of f, which is another such map. This map sends lines and circles to lines and circles, so γ itself must be a straight line or a circle. Moreover, the inverse isometry preserves Euclidean angles, and sends the boundary of the disk back to itself. So γ (or rather, its extension to a full line or full circle) must hit the boundary of the disk at ninety degrees. If γ is a line, then it must go through the origin because these are the only lines which hit the boundary at ninety degrees. So γ is a line through the origin or else a circle hitting the boundary at ninety degrees.

Conversely, all such circles are geodesics. Indeed, let γ be a circle through z_0 hitting the boundary at ninety degrees. There must be a geodesic through z_0 starting in the same direction as γ. This geodesic is a radial line or a circle hitting the boundary at ninety degrees. But a little thought shows that there is only one such curve, so γ itself is a geodesic.



Chapter 3

Extrinsic Theory; Curvature and the Second Fundamental Form

3.1 Differentiating the Normal Vector

The previous chapter was about intrinsic surface theory — the part of the theory that could be understood by a two-dimensional worker living on the surface. We used the surface parameterization $s(u, v)$ to calculate g_{ij}, but our worker would instead measure the g_{ij} directly using a small ruler. All of the remaining mathematics depended only on the g_{ij} without further reference to the surface.

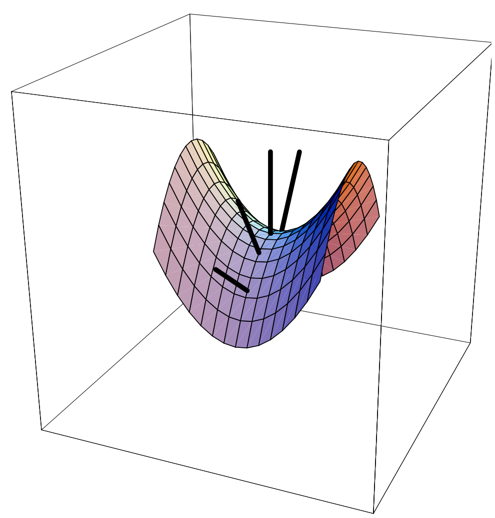
We are now going to study extrinsic surface theory — the part of the theory which requires looking at the surface from the third dimension. Our main goal is to calculate the curvature of the surface. Although we'll continue to calculate in local coordinates, the quantities we study will depend directly on $s(u, v)$; indeed we know from an example in the preface that curvature cannot be determined intrinsically by a two-dimensional worker.

In that preface, we sketched a method to determine the curvature of $\mathcal{S}$ at a point p. Rotate the surface until its tangent plane at p is parallel to the xy-plane; find the equation $z = f(x, y)$ for this rotated surface near p, and expand f in a power series. In practice this method is awkward and we'll use a different technique. Let $\vec{n}$ be a unit normal vector field on the surface. Since $\frac{\partial s}{\partial u}$ and $\frac{\partial s}{\partial v}$ are tangent the surface and linearly independent, their cross product is perpendicular to $\mathcal{S}$ and we can take

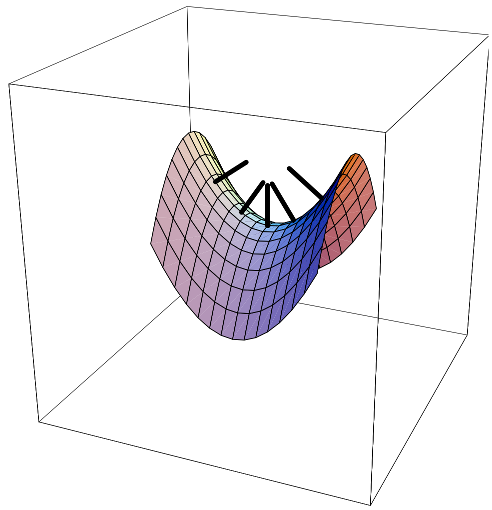
$$\vec{n} = \frac{\frac{\partial s}{\partial u} \times \frac{\partial s}{\partial v}}{\left|\left|\frac{\partial s}{\partial u} \times \frac{\partial s}{\partial v}\right|\right|}$$

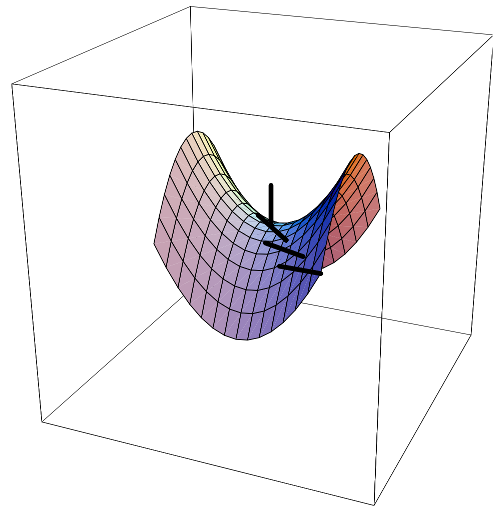
If the surface lies in a plane, this normal vector is constant. Consequently, curvature is related to changes in n; we can study these changes by differentiation. When X is a tangent vector at p, let $X(n)$ be the derivative of n in the direction X. (This derivative will be defined rigorously in the next section.) Since n has length one, $n \cdot n = 1$ and the product rule gives $X(n) \cdot n + n \cdot X(n) = 0$ or $X(n) \cdot n = 0$. Consequently, the derivative $X(n)$ is another vector tangent to the surface.

The picture of the saddle below shows how this works in practice. Pay attention to the origin and move in the x direction. Notice that the change of the normal vectors is also in the positive x-direction. If our saddle has standard form $z = y^2 - x^2$, we will later calculate that the derivative of n in the direction e_1 is $2e_1$.



Repeat this argument, but this time move in the y direction, as in the left picture on the following page. (To avoid interference, the normal vectors have been drawn shorter than they should be.) Notice that the change of the normal vectors is now in the negative y-direction. For a standard saddle, we will later calculate that the derivative of n in the direction e_2 is $-2e_2$.





Finally, move in a diagonal direction X, as in the picture on the right. Notice that the change of the normal vectors is no longer a multiple of X. We will later prove that the derivative of n in the direction $e_1 + e_2$ is $2e_1 - 2e_2$.

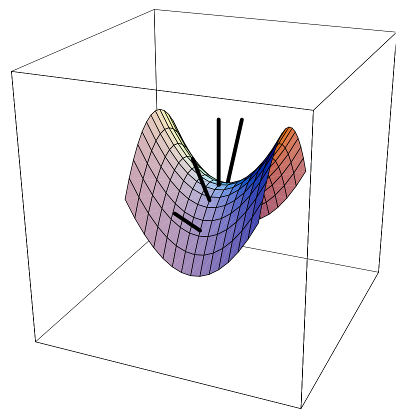
These results are supposed to remind you of eigenvectors. The derivative of n in the direction X is another tangent vector which we will denote $B(X)$. This B is a linear transformation from the tangent space at p to itself. In the example, e_1 and e_2 are eigenvectors of B because B multiples each of these vectors by a scalar. In general, the eigenvalues of B (up to a sign) will be the *principal curvatures* described in the preface.

3.2 Vector Fields in R^3

Recall that a tangent vector at a point p looks like $X = (X_1, X_2)$ in local coordinates. The numbers X_1 and X_2 are *not* the coordinates of the vector in R^3; these are found from the expression

$$X = X_1 \frac{\partial \vec{s}}{\partial u} + X_2 \frac{\partial \vec{s}}{\partial v}$$

We now want to discuss vector fields on a surface $\mathcal{S}$ which point in arbitrary directions in R^3, not just in tangent directions. The normal vector field below is such a field. We will use the letters X, Y, Z to denote tangent vector fields, and the letters U, V, W to denote vector fields in arbitrary directions.



A typical vector field V assigns a vector $V(p) = (V_1, V_2, V_3)$ to each point p in the surface. In local coordinates, p is given by a pair (u, v). Consequently we have

Definition 15 *Let $\mathcal{S}$ be a surface in R^3. A* vector field *on $\mathcal{S}$ is an assignment to each point $p \in \mathcal{S}$ of a three dimensional vector $V(p)$ starting at p, such that the vectors vary from point to point in a C^∞ manner. In local coordinates we have*

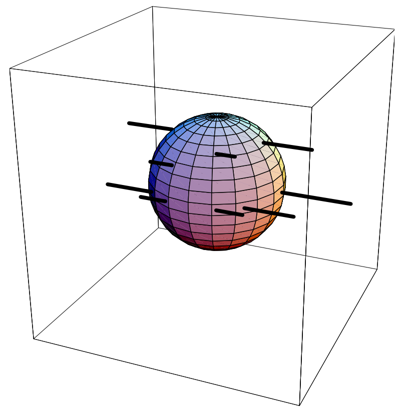
$$V(u, v) = (V_1(u, v), V_2(u, v), V_3(u, v))$$

where the V_i are C^∞ functions.

Notice carefully that when we give a tangent vector (X_1, X_2), the X_i are *not* the three-dimensional coordinates of the vector. But when we give a vector (V_1, V_2, V_3), the V_i *are* just the three-dimensional coordinates of the vector.

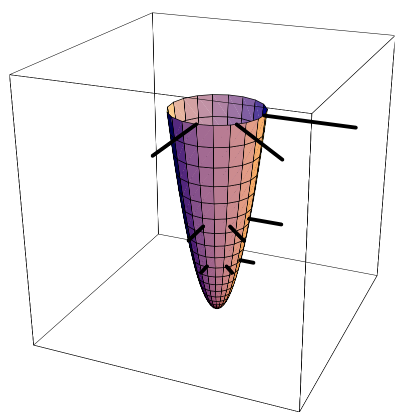
Example 1: Consider the sphere, parameterized in the usual way via spherical coordinates $s(\theta, \phi) = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)$. Let V be the vector field which assigns to each point (x, y, z) the vector $(0, y, 0)$. This vector field is pictured on the next page. Since $y = \sin\phi\sin\theta$, this vector field equals

$$(V_1, V_2, V_3) = (0, \sin\phi\sin\theta, 0)\,.$$



Example 2: Consider the surface $z = x^2 + y^2$. This surface can be parameterized via the map $s(u, v) = (u, v, u^2 + v^2)$. Let V be the vector field which assigns to each point (x, y, z) the vector $z(x, y, 0)$. Then

$$V = (u(u^2 + v^2), v(u^2 + v^2), 0)$$



3.3 Differentiating Vector Fields

We are about to define the derivative of the vector field V in a tangent direction X. The object V must be a *vector field*, and not just a vector at a single point p, since differentiation requires that we compare values of V at nearby points. But X can be a single vector at p, because it merely gives the direction we want to move. However, X must be a *tangent* vector, because if we move in a direction that is not tangent to the surface, we'll leave the surface and V will not longer be defined.

We cannot define the derivative of V in the direction X using the usual definition

$$\frac{d}{dt}V(p+tX)$$

because the line $p+hX$ also leaves the surface and so V is not defined at $p+hX$. But a slight modification works. Let $\alpha(t)$ be a curve on the surface which goes through p at time 0 with direction $\alpha'(0)=X$. Then we can compute

$$\frac{d}{dt}V(\alpha(t))$$

Let us deduce a formula for this derivative. Let $\gamma(t)=(u(t),v(t))$ be a local coordinate representation of α. Then V at $\alpha(t)$ is just

$$(V_1(u(t),v(t)),V_2(u(t),v(t)),V_3(u(t),v(t)))$$

and the desired derivative is

$$\frac{dV(\alpha(t))}{dt}=\left(\frac{\partial V_1}{\partial u}\frac{du}{dt}+\frac{\partial V_1}{\partial v}\frac{dv}{dt},\frac{\partial V_2}{\partial u}\frac{du}{dt}+\frac{\partial V_2}{\partial v}\frac{dv}{dt},\frac{\partial V_3}{\partial u}\frac{du}{dt}+\frac{\partial V_3}{\partial v}\frac{dv}{dt}\right)$$

But the local coordinate expression for $X=\alpha'(0)$ is $(X_1,X_2)=\left(\frac{du}{dt},\frac{dv}{dt}\right)$, so the above expression is

$$\frac{dV(\alpha(t))}{dt}=\left(X_1\frac{\partial V_1}{\partial u}+X_2\frac{\partial V_1}{\partial v},X_1\frac{\partial V_2}{\partial u}+X_2\frac{\partial V_2}{\partial v},X_1\frac{\partial V_3}{\partial u}+X_2\frac{\partial V_3}{\partial v}\right)$$

which is just

$$X_1\frac{\partial\vec{V}}{\partial u}+X_2\frac{\partial\vec{V}}{\partial v}.$$

Notice that this expression depends only on X and not on the particular curve $\alpha(t)$. Consequently,

Definition 16 *Let $X=(X_1,X_2)=X_1\frac{\partial}{\partial u}+X_2\frac{\partial}{\partial v}$ be a tangent vector to a surface $\mathcal{S}$ and let $V=(V_1,V_2,V_3)$ be a three-dimensional vector field. Then*

$$X(V)=X_1\frac{\partial\vec{V}}{\partial u}+X_2\frac{\partial\vec{V}}{\partial v}$$

Example: Let $\mathcal{S}$ be the surface $z=x^2+y^2$ of the previous example and let V be the vector field $z(x,y,0)=(u(u^2+v^2),v(u^2+v^2),0)$ pictured there. Suppose $X=(2,3)$. Then

$$X(V)=2\frac{\partial\vec{V}}{\partial u}+3\frac{\partial\vec{V}}{\partial v}=(6u^2+6uv+2v^2,3u^2+4uv+9v^2).$$

3.4 Basic Differentiation Facts

The rest of our course depends on the straightforward differentiation formula introduced in the previous section. In this section we collect together all of the facts about this differentiation used in the future. These facts are so simple that they are boring. But wait until you see how they are used.

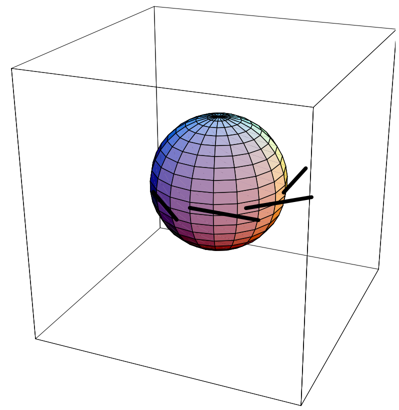
Theorem 23 *Differentiation of a three-dimensional vector field V in the tangent direction X has the following properties*

1. *If r_1 and r_2 are real numbers, $X(r_1 V + r_2 W) = r_1 X(V) + r_2 X(W)$*
2. *If $f(u,v)$ is a function on the surface, $X(f\ V) = X(f)V + fX(V)$*
3. *If r_1 and r_2 are real numbers, $(r_1 X + r_2 Y)(V) = r_1 X(V) + r_2 Y(V)$*
4. *$X\langle V, W\rangle = \langle X(V), W\rangle + \langle V, X(W)\rangle$*
5. *If X and Y are tangent fields, $X(Y) - Y(X) = [X, Y]$*
6. *If X and Y are tangent fields, $X(Y) - Y(X)$ is again a tangent field.*

Proof: Most of these results can be left to the reader. Look back at the fourth result. Since V and W are three-dimensional vectors, the inner product in question is the standard dot product in R^3 rather than the sophisticated two-dimensional dot product defined by the g_{ij}. So this result is just the ordinary product rule

$$X(V_1W_1 + V_2W_2 + V_3W_3) = X(V_1)W_1 + V_1X(W_1) + \ldots = \langle X(V), W\rangle + \langle V, X(W)\rangle$$

The fifth result is our first use of the Lie bracket defined in an earlier section. Its most important consequence is point six. If X and Y are tangent vectors, $X(Y)$ need no longer be tangent to the surface. For example, look at the tangent field shown in the picture below on the sphere, and differentiate it in the direction of the equator; the derivative vectors point inward toward the center of the sphere.



But $X(Y) - Y(X)$ *will be* tangent to the surface, since the Lie bracket of tangent vector fields is again a tangent field.

Let us prove the fifth result. This time Y is a vector field rather than just an isolated vector. In local coordinates $Y = (Y_1(u,v), Y_2(u,v))$. But to differentiate, we must think of this as the three-dimensional vector

$$Y = Y_1(u,v)\frac{\partial s}{\partial u} + Y_2(u,v)\frac{\partial s}{\partial v}$$

Consequently

$$X(Y) = X(Y_1)\frac{\partial s}{\partial u} + Y_1\left(X_1\frac{\partial^2 s}{\partial u^2} + X_2\frac{\partial^2 s}{\partial v \partial u}\right) + X(Y_2)\frac{\partial s}{\partial v} + Y_2\left(X_1\frac{\partial^2 s}{\partial u \partial v} + X_2\frac{\partial^2 s}{\partial v^2}\right)$$

Similarly

$$Y(X) = Y(X_1)\frac{\partial s}{\partial u} + X_1\left(Y_1\frac{\partial^2 s}{\partial u^2} + Y_2\frac{\partial^2 s}{\partial v \partial u}\right) + Y(X_2)\frac{\partial s}{\partial v} + X_2\left(Y_1\frac{\partial^2 s}{\partial u \partial v} + Y_2\frac{\partial^2 s}{\partial v^2}\right)$$

Notice that when we subtract the second expression from the first, the terms involving second partials of s cancel, and we obtain

$$X(Y) - Y(X) = (X(Y_1) - Y(X_1))\frac{\partial s}{\partial u} + (X(Y_2) - Y(X_2))\frac{\partial s}{\partial v}$$

This is a three dimensional tangent vector whose expression in local coordinates is

$$(X(Y_1) - Y(X_1),\ X(Y_2) - Y(X_2))$$

A brief final calculation shows that this is exactly the expression for $[X, Y]$ obtained at the end of section 2.5. QED

3.5 The Normal Field

We apply this theory to the normal vector field $\vec{n}$ consisting of unit vectors perpendicular to the surface. Since $\frac{\partial s}{\partial u}$ and $\frac{\partial s}{\partial v}$ are linearly independent tangent vectors, we have

Definition 17 *Let $\mathcal{S}$ be a surface parameterized by $s(u,v)$. The* normal vector field *is the three-dimensional field*

$$\vec{n} = \frac{\frac{\partial s}{\partial u} \times \frac{\partial s}{\partial v}}{\left|\left|\frac{\partial s}{\partial u} \times \frac{\partial s}{\partial v}\right|\right|}$$

As a matter of fact, there are two choices for this n, differing in sign. By definition, an *orientation* of a surface is a choice of one or the other unit normal. Often this orientation is given geometrically. For instance, closed objects like spheres and doughnuts are usually oriented using outward pointing normals. Graphs like $z = u^2 + v^2$ are usually oriented using upward pointing normals.

If an orientation is given geometrically, it is necessary to check that the previous formula gives the correct choice. When it gives the incorrect orientation, switch u and v to obtain the correct one, or just change of sign of n.

Theorem 24 *If X is a tangent vector, then $X(\vec{n})$ is again tangent to the surface.*

Proof: The normal field has unit vectors, so $\langle n, n \rangle = 1$. Since the derivative of a constant is zero, rule four from section 3.3 gives

$$0 = X \langle n, n \rangle = \langle X(n), n \rangle + \langle n, X(n) \rangle = 2 \langle X(n), n \rangle$$

So $X(n)$ is perpendicular to n, and thus tangent to the surface. QED

Example: Consider the surface $z = x^2 + y^2$. Let us parameterize using polar coordinates, so $s(r,\theta) = \left(r\cos\theta, r\sin\theta, r^2\right)$. Then

$$\frac{\partial s}{\partial r} \times \frac{\partial s}{\partial \theta} = (\cos\theta, \sin\theta, 2r) \times (-r\sin\theta, r\cos\theta, 0) = \left(-2r^2\cos\theta, -2r^2\sin\theta, r\right)$$

Since the length of this vector is $r\sqrt{1+4r^2}$,

$$\vec{n} = \frac{1}{r\sqrt{1+4r^2}}\left(-2r^2\cos\theta, -2r^2\sin\theta, r\right)$$

Let us differentiate $\vec{n}$ in the θ direction. We obtain

$$\frac{1}{r\sqrt{1+4r^2}}\left(2r^2\sin\theta, -2r^2\cos\theta, 0\right)$$

According to the previous theorem, this vector is a tangent vector and thus a linear combination of

$$\frac{\partial s}{\partial r} = (\cos\theta, \sin\theta, 2r)$$

and

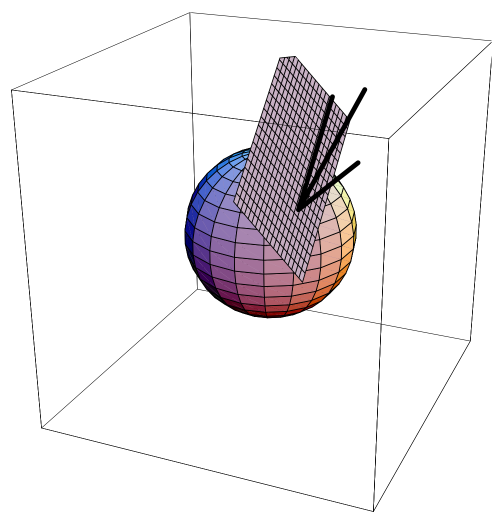
$$\frac{\partial s}{\partial \theta} = (-\sin\theta, \cos\theta, 0)$$

In fact, it is $\frac{-2r}{\sqrt{1+4r^2}}$ times the second vector. We usually write tangent vectors in local coordinates as (X_1, X_2). In this language, the derivative just computed equals

$$\left(0, \frac{-2r}{\sqrt{1+4r^2}}\right)$$

3.6 Decomposing Vectors

In the previous sections, we studied three-dimensional vector fields on a surface $\mathcal{S}$. A vector in this field points off in R^3 willy-nilly, probably at a strange angle to the surface. It is natural to decompose this vector into two pieces, one tangent to the surface and one normal to the surface. This decomposition will lead directly to the central ideas in Gauss' paper.



Theorem 25 *Fix a point $p \in \mathcal{S}$ and let V be a vector at p. Write*

$$\vec{V} = \left\{\vec{V} - \langle V, n\rangle\,\vec{n}\right\} + \langle V, n\rangle\,\vec{n}$$

Then the first term in this expression is tangent to the surface and the second term is normal to the surface.

Proof: The second term is a multiple of n, so it is certainly normal. To show that the first term is tangent to the surface, it suffices to show that the dot product of this term and n is zero. But a glance shows that this is true. QED

Apology: In these notes, I usually omit writing small vectors over V, n, and other expressions. Occasionally I write small vectors to emphasize a point. Thus in the previous theorem, n sometimes has a vector and sometimes does not. It is always the same n.

Remark: It follows from the previous theorem that every three-dimensional vector field V on the surface can be written in the form

$$\vec{V} = \vec{Y} + f\ \vec{n}$$

where Y is tangent to the surface and f is a C^∞ function. Suppose we differentiate this vector field with respect to a tangent vector X. According to the rules in section 3.3, we will obtain

$$X(V) = X(Y) + X(f)n + fX(n)$$

In this expression, $X(f)$ is the directional derivative of the function f in the direction X, which we already understand from the discussion in section 2.3. Consequently, we can understand the derivatives of arbitrary three-dimensional vector fields by studying two special cases:

- $X(Y)$ where X and Y are tangent vector fields
- $X(n)$

3.7 The Fundamental Decomposition

If X and Y are tangent to the surface, the derivative $X(Y)$ may no longer be tangent. See the picture in section 3.4. But the vector $X(Y)$ can be decomposed into a tangential and a normal component. This tangential component is denoted $\nabla_X Y$ and the normal component, which is a multiple of n, is denoted $b(X, Y)n$.

We have already proved that $X(n)$ is tangent to the surface. The resulting tangent vector is denoted $B(X)$.

We have proved the following decomposition result, which is fundamental for everything which follows:

Theorem 26 *If X is a tangent vector and Y is a tangent vector field, we have the following decomposition of derivatives into tangential and normal components:*

$$X(Y) = \nabla_X Y + b(X,Y)\vec{n}$$

$$X(\vec{n}) = B(X)$$

Remark: This chapter is about $b(X,Y)$ and $B(X)$. We will discover that these objects contain exactly the curvature information about the surface, nothing less and nothing more.

The next chapter is about $\nabla_X Y$. According to the current definition, this object can only be calculated by a three-dimensional worker able to differentiate and then separate the resulting vector into a tangential and normal component. However, we will discover that a two-dimensional worker could compute $\nabla_X Y$, using the Christoffel symbols and all that.

Finally, Gauss' *theorema egregium* will arise because the quantities $\nabla_X Y$ and $b(X,Y)$ are not completely independent.

The object $b(X,Y)$ is called *the second fundamental form* on the surface. And yes, the first fundamental form is the metric tensor g_{ij}. More about all that in a moment. Notice carefully that $b(X,Y)$ assigns a *number* to each pair of tangent vectors X and Y, while $B(X)$ assigns a *tangent vector* to each tangent vector. In the next section, we will prove that knowing one of these objects automatically gives the other. We will also prove that

$$B : \text{(tangent space)} \rightarrow \text{(tangent space)}$$

is a linear transformation of a special type — it is symmetric. The linear algebra people have a wonderful theorem about such transformations. We'll describe their theorem in section 3.9 and use it in the rest of the chapter.

3.8 The Crucial Results

Theorem 27 *Let X and Y be tangent vector fields on a surface.*

1. *$b(X, Y)$ and $B(X)$ at a point p depend only on the vectors X and Y at p, and not on their extensions to tangent fields on the surface*
2. *$b(X, Y)$ is linear in X if Y is held fixed, and linear in Y if X is held fixed*
3. *$B(X)$ is linear in X*
4. *$b(X, Y) = b(Y, X)$*
5. *$b(X, Y) = -\langle B(X), Y\rangle$*
6. *$\langle B(X), Y\rangle = \langle X, B(Y)\rangle$*

Proof: The vector $B(X)$ equals $X(n)$, which depends only on X at a point p. The similar result for $b(X, Y)$ then follows once we prove point 5.

Linearity for b and B is a consequence of points one and three of the theorem in section 3.4.

Symmetry of $b(X, Y)$ is proved as follows. From the fundamental decomposition we have

$$X(Y) - Y(X) = \{\nabla_X Y - \nabla_Y X\} + (b(X, Y) - b(Y, X))\vec{n}$$

But $X(Y) - Y(X)$ is tangent to the surface by the theorem in section 3.4, so the normal component of this expression is zero and $b(X, Y) = b(Y, X)$.

Since Y is tangent to the surface and n is normal to the surface, $\langle n, Y\rangle = 0$. Differentiate this expression with respect to X and apply the theorem in section 3.4 to obtain:

$$0 = X\langle n, Y\rangle = \langle X(n), Y\rangle + \langle n, X(Y)\rangle$$

Then apply the decomposition theorem to get

$$0 = \langle B(X), Y\rangle + \langle n, \nabla_X Y + b(X, Y)n\rangle = \langle B(X), Y\rangle + b(X, Y).$$

Here we have used the fact that n and $\nabla_X Y$ are perpendicular and n has length one. So $b(X, Y) = -\langle B(X), Y\rangle$. QED.

3.9 The Principal Axis Theorem

I'd like to remind you of a little linear algebra. Suppose $A = (a_{ij})$ is a matrix. A vector v is an *eigenvector* of A if A just stretches v without rotation, so $Av = \lambda v$. In that case v is an *eigenvalue* of A. There is a mechanical way to find the eigenvalues and eigenvectors of A, summarized below.

For example, consider the matrix $A = \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix}$. Notice that when $v = (1,1)$ we get $Av = 2v$. Notice that when $w = (-1,1)$ we get $Aw = 8w$. So the geometry of A is very simple. It stretches in the v direction by 2 and in the w direction by 8. The numbers 2 and 8 are *secrets of the matrix* which the method of eigenvectors reveals. Neither number is an entry of the matrix.

The grand goal of linear algebra is to reveal the secrets of any matrix by choosing new coordinates which make the matrix easy to understand. Sometimes eigenvalues suffice for this task, and sometimes they are useless. The matrix $B = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ rotates everything by θ and has no eigenvalues at all when $\theta \neq 0, \pi$. The matrix $C = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ has only one eigenvector $v = (1,0)$ up to scalar multiples, so its eigenvectors don't span enough dimensions to tell all of its secrets. The matrix $D = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$ has two eigenvectors $v = (1,0)$ and $w = (1,1)$ with $Dv = 2v$ and $Dw = 3w$, but the eigenvectors are not orthogonal.

There is a beautiful theorem in linear algebra stating a case when eigenvectors tell all of the secrets of A. If our vector space is R^n with its standard dot product, the theorem is

Theorem 28 (The Principal Axis Theorem) *Let A be an $n \times n$ real matrix. Suppose A is symmetric, so that A is unchanged when it is flipped over the diagonal. Then we can find n orthonormal eigenvectors for A. Conversely, if A has n orthonormal eigenvectors, then it is symmetric.*

The symmetry condition is easily seen to be equivalent to the condition that for all v and w we have $Av \cdot w = v \cdot Aw$. This later condition allows us to generalize to an arbitrary vector space:

Theorem 29 (The Principal Axis Theorem) *Let V be a finite dimensional real vector space with an inner product $\langle\ ,\ \rangle$ and let $B : V \to V$ be a linear transformation. Suppose $\langle Bv, w\rangle = \langle v, Bw\rangle$ for all v and w. Then there is an orthonormal basis $e_1, e_2, \ldots, e_n$ for V consisting of eigenvectors of B, so*

$$B(e_i) = \kappa_i\ e_i.$$

We are about to use this theorem when the dimension of the vector space is two. For completeness, we'll recall how to calculate eigenvalues and eigenvectors in this case, and prove the theorem.

Suppose $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a matrix. An *eigenvector* is a nonzero vector $v = (v_1, v_2)$ such that v is merely stretched by A, so $Av = \lambda v$ for some real number λ. This λ is the corresponding *eigenvalue* of A.

If $Av = \lambda v$ for a nonzero v, then $\lambda I - A$ takes the nonzero vector v to zero and conversely. It is easy to check that a 2×2 matrix takes some nonzero vector to zero exactly when its determinant is zero. So the eigenvalues of A are precisely the solutions of

$$P(\lambda) = \det(I\lambda - A) = \det \begin{pmatrix} \lambda - a & -b \\ -c & \lambda - d \end{pmatrix} = 0.$$

For example, the eigenvalues of the matrix $A = \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix}$ discussed earlier are the roots of

$$P(\lambda) = \det \begin{pmatrix} \lambda - 5 & 3 \\ 3 & \lambda - 5 \end{pmatrix} = \lambda^2 - 10\lambda + 16 = (\lambda - 2)\ (\lambda - 8)$$

and so $\lambda = 2, 8$.

Once we know λ, the corresponding eigenvectors are solutions of $Av = \lambda v$, or equivalently $(\lambda I - A)v = 0$. When written out, this will yield two equations, but the equations will be redundant. Thus the solutions form a line through the origin. Any nonzero vector on this line is an eigenvector. For example, in the previous example suppose $\lambda = 2$. Then $(\lambda I - A)v$ becomes

$$\begin{pmatrix} 2-5 & 3 \\ 3 & 2-5 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

which gives the two redundant equations $-3v_1 + 3v_2$ and $3v_1 - 3v_2$. The solutions are vectors with $v_1 = v_2$. One such solution is $v = (1, 1)$, as described earlier.

Finally, we prove the principal axis theorem in the two-dimensional case. The first step is to show that B has at least one eigenvector. Choose an orthonormal basis f_1, f_2 arbitrarily. Write $B(f_1) = af_1 + cf_2$ and $B(f_2) = bf_1 + df_2$ so that the matrix of B will be $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Since the f_i are orthonormal, $\langle Bf_1, f_2 \rangle = c$ and $\langle f_1, Bf_2 \rangle = b$. By hypothesis these are equal, so $b = c$.

The eigenvalue equation is thus

$$P(\lambda) = \det \begin{pmatrix} \lambda - a & -b \\ -b & \lambda - d \end{pmatrix} = \lambda^2 - (a + d)\lambda + (ad - b^2) = 0.$$

The roots of this equation are

$$\lambda = \frac{(a + d) \pm \sqrt{(a + d)^2 - 4(ad - b^2)}}{2} = \frac{(a + d) \pm \sqrt{(a - d)^2 + 4b^2}}{2}.$$

Since the expression under the square root sign is nonnegative, this equation has real roots. So at least one eigenvector exists.

Let e_1 be an eigenvector. We may multiply e_1 by a constant, so assume it has length one. Choose a perpendicular vector e_2 of length one. Then by hypothesis $\langle e_1, Be_2\rangle = \langle Be_1, e_2\rangle = \langle \kappa_1 e_1, e_2\rangle = 0$. Hence Be_2 is perpendicular to e_1 and therefore must be a multiple of e_2. So e_2 is also an eigenvector of B. QED.

3.10 Principal Curvatures

We can now put everything together. Consider the linear transformation

$$B : (\text{tangent space at } p) \to (\text{tangent space at } p).$$

This B is a symmetric map from the two-dimensional tangent space to itself. By the principal axis theorem, it has an orthonormal basis of eigenvectors.

Definition 18 *Let $\mathcal{S}$ be a surface with a fixed orientation, $p \in \mathcal{S}$.*

1. *The eigenvalues of $-B$ are denoted κ_1 and κ_2 and called the* principal curvatures *of the surface at p*
2. *The corresponding eigenvectors are called the* principal directions *at p. These directions are defined unless $\kappa_1 = \kappa_2$. When $\kappa_1 = \kappa_2$, all vectors are eigenvectors and the principal directions are not well-defined.*
3. *The product $\kappa = \kappa_1\ \kappa_2$ is called the* Gaussian curvature *of the surface at p*
4. *The sum $m = \kappa_1 + \kappa_2$ is called the* mean curvature *of the surface at p.*

Remark: We choose the eigenvalues of $-B$ rather than B to follow the historical convention. In the rest of this chapter, we will explain how to compute b, B, κ_1, and κ_2. In particular, we will justify these definitions by showing that they agree with the numbers defined in the preface.

3.11 A Formula for b

We are going to find a formula for the second fundamental form $b(X,Y)$. Write X in coordinates as $X = (X_1, X_2) = X_1\frac{\partial}{\partial u} + X_2\frac{\partial}{\partial v} = \sum X_i \frac{\partial}{\partial u_i}$. Similarly $Y = \sum Y_j \frac{\partial}{\partial u_j}$. Since b is bilinear,

$$b(X,Y) = \sum_{ij} b\left(\frac{\partial}{\partial u_i}, \frac{\partial}{\partial u_j}\right) X_i Y_j = \sum_{ij} b_{ij}\ X_i Y_j$$

where $b_{ij} = b\left(\frac{\partial}{\partial u_i}, \frac{\partial}{\partial u_j}\right)$. In language which used to be popular years ago, the b_{ij} is a *tensor of rank two.* When this formula is written out in detail, we get an expression similar to our earlier formula for ds^2 in terms of the g_{ij}, namely

$$b(X,Y) = b_{11}X_1Y_1 + b_{12}\left(X_1Y_2 + X_2Y_1\right) + b_{22}X_2Y_2.$$

It remains to compute $b_{ij} = b\left(\frac{\partial}{\partial u_i}, \frac{\partial}{\partial u_j}\right)$. Recall that $b(X,Y)$ is the normal component of the vector $X(Y)$. In our case, $Y = \frac{\partial}{\partial u_j}$, which corresponds to the three-dimensional vector $\frac{\partial s}{\partial u_j}$. The derivative of this vector with respect to $X = \frac{\partial}{\partial u_i}$ is then $\frac{\partial^2 s}{\partial u_i \partial u_j}$. This vector need not be normal, but its normal component is

$$b_{ij} = \frac{\partial^2 \vec{s}}{\partial u_i \partial u_j} \cdot \vec{n}$$

We have proved

Theorem 30 *The second fundamental form $b(X,Y)$ is given by*

$$b(X,Y) = b\left(\sum_i X_i \frac{\partial}{\partial u_i},\ \sum_j Y_j \frac{\partial}{\partial u_j}\right) = \sum_{ij} b_{ij}\ X_i Y_j$$

where $b = (b_{ij})$ is the matrix with entries

$$b_{ij} = \frac{\partial^2 \vec{s}}{\partial u_i \partial u_j} \cdot \vec{n}$$

Remark: We can write this result more concretely if our surface has the form $z = f(x,y)$ so that $s(x,y) = (x, y, f(x,y))$. Then it is useful to write $\frac{\partial f}{\partial x} = f_x$, etc., and we have

$$\frac{\partial s}{\partial x} \times \frac{\partial s}{\partial y} = (1, 0, f_x) \times (0, 1, f_y) = (-f_x, -f_y, 1)$$

Moreover, $\frac{\partial^2 s}{\partial x^2} = (0, 0, f_{xx})$, with similar results for other second partials. The dot product $\frac{\partial^2 s}{\partial x^2} \cdot \vec{n}$ is then f_{xx} divided by the length of $(-f_x, -f_y, 1)$, and we ultimately obtain

Theorem 31 *If a surface is given by the equation $z = f(x,y)$, then b_{ij} is the matrix*

$$b = \frac{1}{\sqrt{1 + f_x^2 + f_y^2}} \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix}$$

3.12 A Formula for B

The map B is a linear transformation from the set of tangent vectors to itself. In coordinates

$$B(X) = B(X_1, X_2) = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

where B_{ij} are the entries of the matrix B. We easily deduce that

$$B\left(\frac{\partial}{\partial u_i}\right) = \sum_j B_{ji} \frac{\partial}{\partial j}.$$

Now apply the formula $\langle B(X), Y\rangle = -b(X,Y)$ to $X = \frac{\partial}{\partial u_i}$ and $Y = \frac{\partial}{\partial u_k}$. We obtain

$$\left\langle B\left(\frac{\partial}{\partial u_i}\right), \frac{\partial}{\partial u_k}\right\rangle = -b\left(\frac{\partial}{\partial u_i}, \frac{\partial}{\partial u_k}\right) = -b_{ik}$$

or

$$\left\langle \sum_j B_{ji} \frac{\partial}{\partial j}, \frac{\partial}{\partial u_k}\right\rangle = -b_{ik}.$$

The expression on the left is $\sum B_{ji} g_{jk}$, but it is better to recall that $g_{ij} = g_{ji}$ and $b_{ik} = b_{ki}$ and write our equation in the form

$$\sum_j g_{kj} B_{ji} = -b_{ki}.$$

We have proved

Theorem 32 *The matrix B satisfies the equation*

$$g\ B = -b$$

and consequently can be computed using the formula

$$B = -(g^{-1})\ b$$

Remark: We obtain a more concrete formula when $z = f(x,y)$. Then $s(x,y) = (x, y, f(x,y))$ and so $\frac{\partial s}{\partial x} = (1, 0, f_x)$ and $\frac{\partial s}{\partial y} = (0, 1, f_y)$. Therefore $g_{11} = 1 + f_x^2, g_{12} = f_x f_y$, and $g_{22} = 1 + f_y^2$. The determinant of the matrix g is then

$$(1 + f_x^2)\ (1 + f_y^2) - (f_x f_y)^2 = 1 + f_x^2 + f_y^2$$

and so

$$g^{-1} = \frac{1}{1 + f_x^2 + f_y^2} \begin{pmatrix} 1 + f_y^2 & -f_x f_y \\ -f_x f_y & 1 + f_x^2 \end{pmatrix}.$$

Consequently,

Theorem 33 *If $z = f(x, y)$, the matrix B is given by the formula*

$$B = -\frac{1}{(1 + f_x^2 + f_y^2)^{3/2}} \begin{pmatrix} 1 + f_y^2 & -f_x f_y \\ -f_x f_y & 1 + f_x^2 \end{pmatrix} \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix}$$

3.13 Justification of the Definition

Suppose $\mathcal{S}$ is a surface containing the point p and we wish to compute the principal curvatures at p. According to the definition, these curvatures are the negatives of the eigenvalues of the matrix B defined by $B(X) = X(n)$. Thus we can find linearly independent tangent vectors e_1 and e_2 at p which satisfy $e_1(n) = -\kappa_1 e_1$ and $e_2(n) = -\kappa_2 e_2$.

Suppose we were to rotate the surface in three space. We claim that the principle curvatures will not change. Indeed, n will rotate and the e_i will rotate and the *derivative contraption* will rotate, and after this rotation we will still have $e_1(n) = -\kappa_1 e_1$ and $e_2(n) = -\kappa_2 e_2$ at the rotated image of p.

Clearly we can rotate the surface so p lies over the origin and its tangent plane is parallel to the xy-plane. After this rotation, the surface will be given near p by

$$z = f(x, y) = f(0, 0) + \frac{1}{2}\left(ax^2 + 2bxy + cy^2\right) + \ldots$$

Since the tangent plane at the origin is parallel to the xy-plane, we have $f_x = f_y = 0$ at the origin, and $f_{xx} = a, f_{xy} = b, f_{yy} = c$. Therefore the matrix B at the origin is

$$-\frac{1}{(1 + 0 + 0)^{3/2}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} = \begin{pmatrix} -a & -b \\ -b & -c \end{pmatrix}$$

The principal curvatures are the negatives of the eigenvalues of this B. So the principal curvatures and principal directions are the eigenvalues and eigenvectors of

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

In the special case when $b = 0$, these eigenvalues are a and c and the principal directions are along the x and y axes. Therefore we recover the description of curvature from the preface:

$$f(x, y) = f(0, 0) + \frac{\kappa_1}{2}x^2 + \frac{\kappa_2}{2}y^2 + \ldots$$

In the general case when b may not be zero, write $p = (x, y)$ and notice that

$$\langle B(p), p\rangle = \left\langle \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix} \right\rangle = ax^2 + 2bxy + cy^2$$

This is the quadratic term of the Taylor expansion, up to a factor of 1/2. Choose orthonormal eigenvectors e_1 and e_2. Then $p = (x, y)$ can be written in the form $ue_1 + ve_2$, where u and v are new coordinates of p in the coordinate system defined by e_1 and e_2. We have

$$ax^2 + 2bxy + cy^2 = \langle B(p), p\rangle = \langle B(ue_1 + ve_2), ue_1 + ve_2\rangle .$$

Since the e_i are eigenvectors, this equals

$$\left\langle u\kappa_1 e_1 + v\kappa_2 e^2, ue_1 + ve_2\right\rangle = \kappa_1 u^2 + \kappa_2 v^2.$$

Hence in the new coordinates, our surface has the equation

$$f(u, v) = f(0, 0) + \frac{\kappa_1}{2}u^2 + \frac{\kappa_2}{2}v^2 + \ldots$$

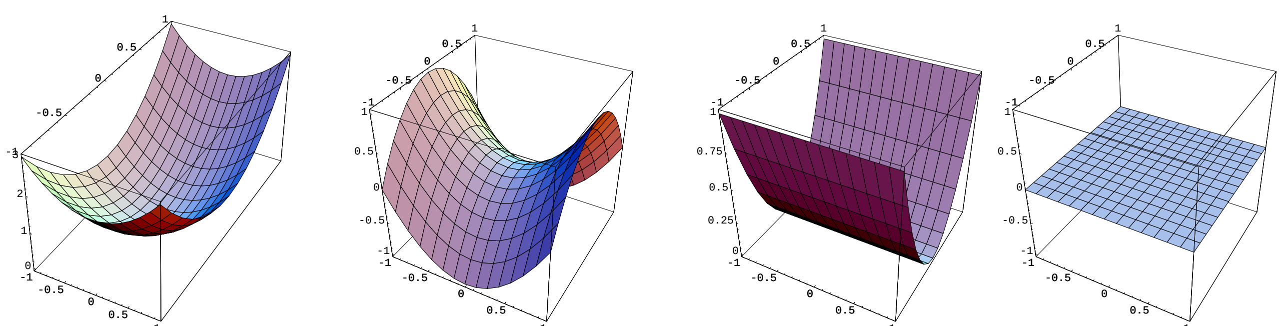
We have proved

Theorem 34 *Let $\mathcal{S}$ be a surface containing a point p. Rotate the surface until p lies above the origin, the tangent plane to the surface at p is parallel to the xy-plane, and the principal directions at p point along the x and y axes. Then the equation of the rotated surface has the form*

$$f(x, y) = f(0, 0) + \frac{\kappa_1}{2}x^2 + \frac{\kappa_2}{2}y^2 + \ldots$$

where κ_1 and κ_2 are the principal curvatures.

Remark: It is best to think of this theorem pictorially. It states that up to second order terms, our surface looks like one of the pictures below, provided we orient our surface along the principal directions.



3.14 Gaussian Curvature and Mean Curvature

If κ_1 and κ_2 are the principal curvatures, recall that we have defined the *mean curvature* $m = \kappa_1 + \kappa_2$ and the *Gaussian curvature* $\kappa = \kappa_1 \kappa_2$. Since the principal curvatures are the negatives of the eigenvalues of B, we have

$$\det(\lambda I - B) = (\lambda + \kappa_1)\ (\lambda + \kappa_2) = \lambda^2 + m\lambda + \kappa.$$

However, a brief calculation shows that for any matrix B,

$$\det(\lambda I - B) = \lambda^2 - \mathrm{tr}(B)\lambda + \det(B)$$

where $\mathrm{tr}(B)$, the *trace* of B, is the sum of the diagonal elements of B, and $\det(B)$ is the determinant of B. Consequently, we have proved

Theorem 35 *The mean curvature and Gaussian curvature are given by*

$$m = -tr(B) \qquad\qquad \kappa = \det(B).$$

In particular, we can apply our previous calculations to deduce

Theorem 36 *If $\mathcal{S}$ is given by $s(u,v)$, then*

$$\kappa = \det(g^{-1})\det(b) = \frac{\det(b)}{\det g} = \frac{\left(\frac{\partial^2 s}{\partial u^2} \cdot n\right)\ \left(\frac{\partial^2 s}{\partial v^2} \cdot n\right) - \left(\frac{\partial^2 s}{\partial u \partial v} \cdot n\right)^2}{\left(\frac{\partial s}{\partial u} \cdot \frac{\partial s}{\partial u}\right)\ \left(\frac{\partial s}{\partial v} \cdot \frac{\partial s}{\partial v}\right) - \left(\frac{\partial s}{\partial u} \cdot \frac{\partial s}{\partial v}\right)^2}$$

If $\mathcal{S}$ is given by $z = f(x,y)$, then

$$\kappa = \frac{f_{xx} f_{yy} - (f_{xy})^2}{\left(1 + f_x^2 + f_y^2\right)^2}$$

Remark: If B is an arbitrary linear transformation, the numbers $\mathrm{tr}(B)$ and $\det(B)$ are invariants which do not depend on the choice of basis. It can be proved that they are the only such invariants in the sense that any continuous invariant is just a function of $\mathrm{tr}(B)$ and $\det(B)$. Therefore from an algebraic point of view, the Gaussian curvature and the mean curvature are natural invariants of the operator B.

As explained in the preface, we will soon prove that the Gaussian curvature can be computed intrinsically by a two-dimensional worker. It will interest us greatly. The mean curvature is also interesting, although we do not have time to study it in this course. For instance, if a wire is bent and dipped into a soap film, a surface forms supported by the wire. The chosen surface has minimal area among all surfaces bounded by the wire, and it is then easy to prove that it has mean curvature zero using the variational techniques introduced during the discussion of geodesics. So soap films bounding a wire are always shaped like a saddle. The curvatures of this saddle vary from point to point, but they are always of equal magnitude and opposite sign.

3.15 Examples

Consider a surface of revolution obtained by revolving $y = f(x)$ around the x-axis. In section 2.12, we parameterized such a surface by $s(x, \theta) = (x, f(x)\cos\theta, f(x)\sin\theta)$ and obtained

$$\frac{\partial s}{\partial x} = (1, f'\cos\theta, f'\sin\theta)$$

$$\frac{\partial s}{\partial \theta} = (0, -f\sin\theta, f\cos\theta)$$

$$g_{11} = 1 + (f'(x))^2 \qquad g_{12} = 0 \qquad g_{22} = (f(x))^2$$

A brief calculation gives

$$n = \frac{1}{\sqrt{1 + (f'(x))^2}} \big(f'(x), -\cos\theta, -\sin\theta\big)$$

Notice that this normal points inward. Let us change the sign of the normal and use the more standard outward-pointing normal.

We have

$$\frac{\partial^2 s}{\partial x^2} = (0, f''\cos\theta, f''\sin\theta)$$

$$\frac{\partial^2 s}{\partial x \partial \theta} = (0, -f'\sin\theta, f'\cos\theta)$$

$$\frac{\partial^2 s}{\partial \theta^2} = (0, -f\cos\theta, -f\sin\theta)$$

The matrix for b is obtained by dotting these terms with n; recall that we have changed the sign of n.

$$b = \frac{1}{\sqrt{1 + (f')^2}} \begin{pmatrix} f'' & 0 \\ 0 & -f \end{pmatrix}$$

Consequently

$$B = -g^{-1}b = -\begin{pmatrix} \frac{1}{1+(f')^2} & 0 \\ 0 & \frac{1}{f^2} \end{pmatrix} \frac{1}{\sqrt{1 + (f')^2}} \begin{pmatrix} f'' & 0 \\ 0 & -f \end{pmatrix}$$

and so

$$B = \begin{pmatrix} \frac{-f''}{(1+(f')^2)^{3/2}} & 0 \\ 0 & \frac{1}{f(1+(f')^2)^{1/2}} \end{pmatrix}$$

This matrix is already diagonal, and the principal curvatures are the negatives of its diagonal entries. Therefore the principal directions are along the axis with varying x and constant θ, giving curvature

$$\kappa_1 = \frac{f''}{(1+(f')^2)^{3/2}}$$

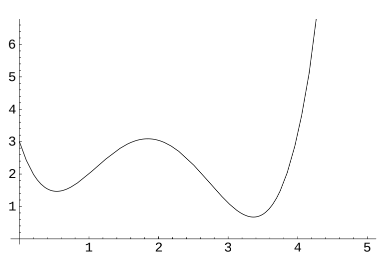
and along the meridians with varying θ and constant x, giving curvature

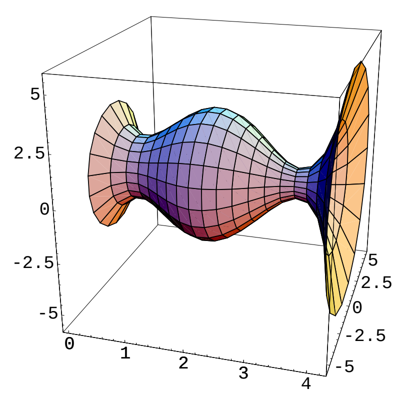
$$\kappa_2 = \frac{-1}{f(1+(f')^2)^{1/2}}.$$

The Gaussian curvature is the product of these numbers,

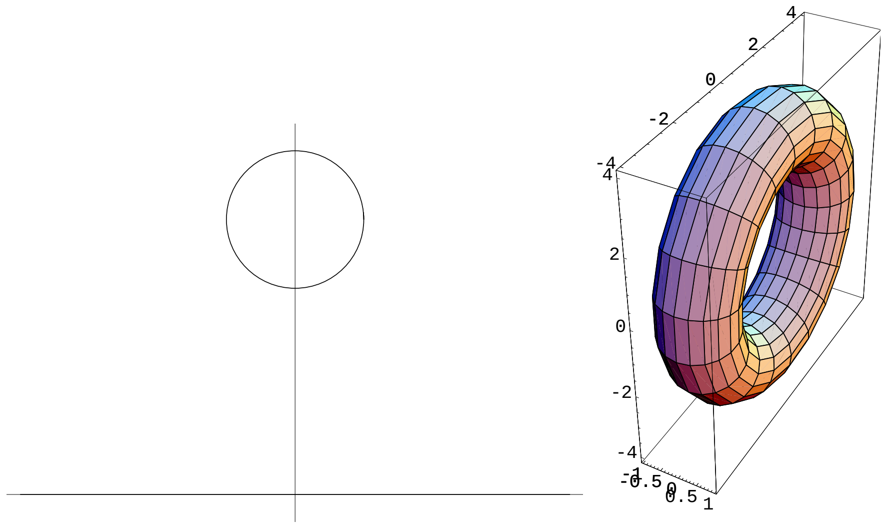
$$\kappa = \frac{-f''}{f(1+(f')^2)^2}$$

Remark: A little thought shows that these results are reasonable. Consider the picture below. Notice that when f is concave up so that $f'' > 0$, the two curvatures have opposite signs and the surface of revolution looks like a saddle. But when f is concave down so that $f'' < 0$, the two curvatures have the same sign and the surface looks like a paraboloid.

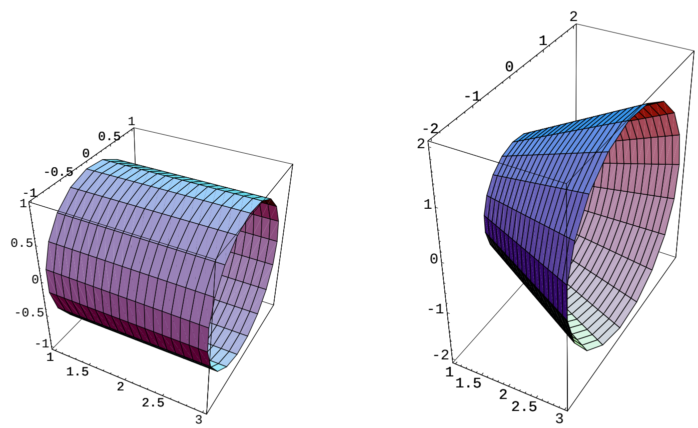




Since we can obtain a doughnut by rotating a circle, these results imply that the doughnut looks like a saddle along the inside half, and like a paraboloid along the outside half.

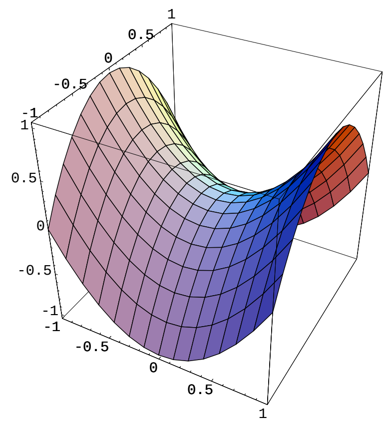


An interesting special case occurs when $f''(x) = 0$ so that $f(x) = ax + b$. In this case $\kappa_1 = 0$ and the Gaussian curvature is zero. The corresponding surfaces look like cylinders or cones.



Consider the case $f(x) = \sqrt{a^2 - x^2}$, which yields a sphere of radius a. A brief calculation shows that $\kappa_1 = \kappa_2 = -\frac{1}{a}$ and $\kappa = \frac{1}{a^2}$.

Let us finally study the saddle $y = \frac{y^2}{2} - \frac{x^2}{2}$ at arbitrary points rather than just at the origin.



By theorem 33, the matrix B for this surface is

$$B = \frac{-1}{(1 + x^2 + y^2)^{3/2}} \begin{pmatrix} 1 + y^2 & xy \\ xy & 1 + x^2 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

This matrix equals

$$B = \frac{1}{(1 + x^2 + y^2)^{3/2}} \begin{pmatrix} 1 + y^2 & -xy \\ xy & -1 - x^2 \end{pmatrix}$$

and so $\mathrm{tr}(B) = \frac{(y^2 - x^2)}{(1+x^2+y^2)^{3/2}}$ and $\det(B) = \frac{-1-y^2-x^2}{(1+x^2+y^2)^3}$. The characteristic polynomial for $-B$ is then $\lambda^2 + \mathrm{tr}(B)\lambda + \det(B) = \lambda^2 + \frac{y^2-x^2}{(1+x^2+y^2)^{3/2}}\lambda + \frac{-1-y^2-x^2}{(1+x^2+y^2)^3}$ and the eigenvalues of this polynomial are the principal curvatures

$$\kappa_i = \frac{(x^2 - y^2) \pm \sqrt{(x^2 - y^2)^2 + 4(1 + x^2 + y^2)}}{2(1 + x^2 + y^2)^{3/2}}$$

At the origin, these values are ± 1, confirming earlier results. Because the expression $1 + x^2 + y^2$ inside the square root is always positive, one of these terms is positive and one is negative. The two curvatures approach zero when x and y are large.

3.16 Algebraic Postscript

The theory of b and B developed above can be formulated purely algebraically. From this standpoint, differential geometry enters the picture only in the initial definition $b(X,Y) = X(Y){\cdot}n$. In linear algebra courses, the theory is known as the *theory of quadratic forms*.

Here is a sketch. Let V be a finite dimensional real vector space. A quadratic form on V is a symmetric bilinear map $b(X,Y)$ from pairs of vectors X and Y to the real numbers. The associated quadratic map is the map $b(X,X)$. In coordinates, this quadratic map has the form

$$(r_1,\ldots,r_n) \to \sum b_{ij} r_i r_j$$

Knowing $b(X,Y)$ determines the quadratic map, but conversely the quadratic map determines b because $b(X,Y) = \frac{1}{2}\left(b(X+Y,X+Y) - b(X,X) - b(Y,Y)\right)$. The fundamental theorem of quadratic form theory states that we can always choose new coordinates so that the quadratic map becomes

$$(s_1,\ldots,s_n) \to \pm s_1^2 + \ldots \pm s_k^2$$

Moreover, the number of terms with a plus sign, the number with a minus sign, and the number which do not appear at all, are invariants of b.

Now suppose that V has an inner product. A deeper theory emerges if we pay attention to this inner product. Then the fundamental theorem states that we can always choose a new *orthonormal* basis so the quadratic map becomes

$$(s_1,\ldots,s_n) \to \kappa_1 s_1^2 + \ldots + \kappa_n s_n^2$$

and the κ_i are invariants of b.

The first step of the proof is to construct a linear transformation $B : V \to V$ so $b(X,Y) = \langle B(X), Y\rangle$. Once this has been done, the fundamental theorem follows from (and is equivalent to) the principal axis theorem. In our course, the linear transformation B appears naturally from the geometry. Notice that we used $-B$ instead of B for historical reasons.

Chapter 4

The Covariant Derivative

4.1 Introduction

Euclid's geometry book has no introduction. Instead the book starts with the following definitions:

1. A **point** is that which has no part.
2. A **line** is breadthless length.
3. The extremities of a line are points.
4. A **straight line** is a line that lies evenly with the points on itself.
5. A **surface** is that which has length and breadth only.
6. The extremities of a surface are lines.
7. A **plane surface** is a surface which lies evenly with the straight lines on itself.

This first page of Euclid is almost a summary of our course. Euclid's *lines* are our *curves*. His *straight lines* are our *geodesics*. His *surfaces* are our *surfaces*, and his *plane surfaces* are our *zero curvature surfaces*.

Let us examine carefully the implications of the statement that *a two-dimensional worker will think that geodesics are straight lines.*

Suppose a two-dimensional worker is living on a surface and examines a path $\gamma(t)$ starting in the direction $\gamma'(0)$. The worker can construct the geodesic $g(t)$ which starts in the same direction $g'(0) = \gamma'(0)$. To this worker, γ will *curve* if it bends away from this geodesic.

Curvature as defined by such a two-dimensional worker is called *geodesic curvature* and denoted κ_g.

In this chapter, we will learn how to compute the geodesic curvature. It turns out to be easiest to develop the theory by first learning how to differentiate vector fields.

4.2 Review

If X and Y are tangent vector fields, the derivative $X(Y)$ need not be tangent to the surface. In the previous chapter, we decomposed $X(Y)$ as

$$X(Y) = \nabla_X Y + b(X,Y)\vec{n}.$$

We want to obtain a formula for $\nabla_X Y$. We will discover that $\nabla_X Y$ can be computed by a two-dimensional worker.

Our final formula will be very beautiful. Suppose $X = (1,0) = \frac{\partial}{\partial u}$. Then we will prove that

$$\nabla_{\frac{\partial}{\partial u}} (Y_1, Y_2) = \left(\frac{\partial Y_1}{\partial u} + \Gamma_{11}^1 Y_1 + \Gamma_{12}^1 Y_2, \frac{\partial Y_2}{\partial u} + \Gamma_{11}^2 Y_1 + \Gamma_{12}^2 Y_2 \right)$$

And yes, those are the Christoffel symbols from our study of geodesics. A similar result holds for $\frac{\partial}{\partial v}$ and the general result is a linear combination of these two special results.

You might suspect that the following easier formula should hold. But we'll explain in the next section why this formula couldn't possibly be correct.

$$\frac{\partial}{\partial u}(Y_1, Y_2) = \left(\frac{\partial Y_1}{\partial u}, \frac{\partial Y_2}{\partial u} \right)$$

4.3 The Debauch of Indices

In the early part of this century, differential geometry was intimately connected to *tensor calculus*, an unpleasant subject involving manipulation of complicated multi-indexed symbols. In tensor analysis, we associate indexed symbols with the local coordinate versions of geometric objects on the surface. New objects are then created using various algebraic combinations of derivatives of the original symbols. These new expressions can be complicated, leading to what Spivak called "the debauch of indices."

But there is a catch. If we are not careful, the new objects will depend on the coordinate system chosen.

I'd like to show you an example. Suppose X and Y are tangent vector fields. In local coordinates $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$, so X and Y are tensors of a simple type. Suppose we want to define the derivative of Y in the direction X. It is natural to mimic the definition in section 3.3 and write

$$X(Y) = X_1 \frac{\partial Y}{\partial u} + X_2 \frac{\partial Y}{\partial v}$$

This is exactly the formula we used to define $X(V)$ except that V was a three-dimensional vector in our earlier work. But there is an extremely importance difference. We deduced our earlier formula from an invariant expression

$$\frac{d}{dt} V(\alpha(t))$$

that makes direct sense on the surface. In the present case, we have written down an expression algebraically without providing a corresponding surface calculation.

It turns out that the proposed definition of $X(Y)$ is nonsense because $X(Y)$ gives different results in different local coordinate systems.

Here is an example. Consider the surface $s(x, y) = (x, y, 0)$. It is just the xy-plane. Let X be the local coordinate vector $(1, 0)$ and let Y be the local coordinate vector $(-y, x)$. Our definition of $X(Y)$ gives $\frac{d}{dx}(-y, x) = (0, 1)$. In particular, this derivative is not zero.

However, watch what happens when we do the same calculation in polar coordinates $s(r, \theta) = (r \cos\theta, r \sin\theta, 0)$. Then the local coordinate vector $(0, 1)$ corresponds to the three-dimensional vector $\frac{\partial s}{\partial \theta} = (-r \sin\theta, r \cos\theta, 0) = (-y, x, 0)$. This is exactly the Y we used earlier. But in polar coordinates $X(Y)$ will be zero regardless of the coordinate expression for X because $Y = (0, 1)$ and the coefficients of Y are constant.

4.4 Fundamental Properties

In section 3.4, we proved some very straightforward properties of vector differentiation $X(V)$. These properties lead to a list of similar properties for $\nabla_X Y$, which are given in the next theorem. Surprisingly, these properties completely determine $\nabla_X Y$, even though the formula for $\nabla_X Y$ is considerably more complex than we might have imagined.

Theorem 37 *Let X be a tangent vector at p, and let Y be a tangent vector field. Then vector differentiation $\nabla_X Y$ has the following properties:*

1. $\nabla_X(r_1 Y + r_2 Z) = r_1 \nabla_X Y + r_2 \nabla_X Z$
2. *If f is a C^∞ function,* $\nabla_X fY = X(f)\ Y + f\ \nabla_X Y$
3. $\nabla_{r_1 X + r_2 Y} Z = r_1 \nabla_X Z + r_2 \nabla_Y Z$
4. $X \langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X Z \rangle$
5. *If X and Y are tangent vector fields,* $\nabla_X Y - \nabla_Y X = [X, Y]$

Proof: By the theorem in section 3.4, vector differentiation is linear. So $X(r_1 Y + r_2 Z) = r_1 X(Y) + r_2 X(Z)$. Take the tangential component of both sides of this equation. The component on the left is $\nabla_X(r_1 Y + r_2 Z)$ and the component on the right is $r_1 \nabla_X Y + r_2 \nabla_X Z$. Part three is proved the same way.

By the same theorem,

$$X(fY) = X(f)Y + fX(Y)$$

Decomposing this equation into components gives

$$\nabla_X(fY) + b(X, fY)\vec{n} = \{X(f)Y + f\nabla_X Y\} + fb(X, Y)\vec{n}$$

and the third result follows by taking the tangential component of both sides.

According to the earlier theorem, $X \langle Y, Z \rangle$ equals

$$\langle X(Y), Z \rangle + \langle Y, X(Z) \rangle = \langle \nabla_X Y + b(X, Y)\vec{n}, Z \rangle + \langle Y, \nabla_X Z + b(X, Z)\vec{n} \rangle$$

But Z and n are orthogonal since Z is a tangent vector. Similarly Y and n are orthogonal. Therefore the above result becomes

$$X \langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X Z \rangle$$

Finally we prove the last result. According to the earlier theorem, $X(Y) - Y(X) = [X, Y]$. Decomposing into tangential and normal components gives

$$\nabla_X Y + b(X, Y)\vec{n} - \nabla_Y X - b(Y, X)\vec{n} = [X, Y]$$

and the result follows by equating the tangential components of both sides.

4.5 A Formula for $\nabla_X Y$

Suppose $X = \sum X_i \frac{\partial}{\partial u_i}$ is a vector and $Y = \sum Y_j(u,v) \frac{\partial}{\partial u_j}$ is a vector field. Applying the previous theorem,

$$\nabla_X Y = \sum_{ij} X_i \left\{ \nabla_{\frac{\partial}{\partial u_i}} \left(Y_j \frac{\partial}{\partial u_j} \right) \right\}$$

By the second part of the theorem in section 4.3, we have

$$\nabla_{\frac{\partial}{\partial u_i}} \left(Y \frac{\partial}{\partial u_j} \right) = \left(\frac{\partial Y}{\partial u_i} \right) \frac{\partial}{\partial u_j} + Y \left(\nabla_{\frac{\partial}{\partial u_i}} \frac{\partial}{\partial u_j} \right)$$

So

$$\nabla_X Y = \sum_{ij} X_i \left\{ \frac{\partial Y_j}{\partial u_i} \frac{\partial}{\partial u_j} + Y_j \nabla_{\frac{\partial}{\partial u_i}} \frac{\partial}{\partial u_j} \right\}$$

Remark: We now need a formula for $\nabla_{\frac{\partial}{\partial u_i}} \frac{\partial}{\partial u_j}$. This expression is another vector field, and so a linear combination of the $\frac{\partial}{\partial u_k}$. It turns out that the coefficients of this linear expression are the Christoffel symbols. Let us forget for a moment that we have worked with these symbols earlier and give

Definition 19 *The Christoffel symbols Γ_{ij}^k are defined by the formula*

$$\nabla_{\frac{\partial}{\partial u_i}} \frac{\partial}{\partial u_j} = \sum_k \Gamma_{ij}^k \frac{\partial}{\partial u_k}.$$

Example: When $i = j = 1$, this formula states

$$\nabla_{\frac{\partial}{\partial u}} \frac{\partial}{\partial u} = \Gamma_{11}^1 \frac{\partial}{\partial u} + \Gamma_{11}^2 \frac{\partial}{\partial v}.$$

Rewriting the final expression slightly, we obtain

Theorem 38 (The Covariant Differentiation Formula) *Suppose $X = \sum X_i \frac{\partial}{\partial u_i}$ and $Y = \sum Y_i \frac{\partial}{\partial u_i}$. Then*

$$\nabla_X Y = \sum_{ik} X_i \left\{ \frac{\partial Y_k}{\partial u_i} + \sum_j \Gamma_{ij}^k Y_j \right\} \frac{\partial}{\partial u_k}$$

Remark: In this expression, the outer sum gives the derivative as a linear combination of the basis vectors. The above formula is more palatable if we look at the special case when $X = \frac{\partial}{\partial u}$. According to the formula, the derivative of (Y_1, Y_2) in this direction is

$$\sum_k \left\{ \frac{\partial Y_k}{\partial u} + \sum_k \Gamma^k_{1j} Y_j \right\} \frac{\partial}{\partial u_k}$$

The first term inside the curley bracket is the naive term we wrote down in section 4.2. The second term is a correction which yields a coordinate invariant derivative.

To finish the theory, we must find a formula for Γ^k_{ij}. Luckily, the correct formula is the formula we obtained in chapter three:

Theorem 39 *The Γ^k_{ij} are given by the following formula. Consequently, they are exactly equal to the Christoffel symbols which occurred earlier in our course.*

$$\Gamma^l_{ij} = \frac{1}{2} \sum_k (g^{-1})_{lk} \left\{ \frac{\partial g_{jk}}{\partial u_i} + \frac{\partial g_{ik}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_k} \right\}$$

Proof: By the fourth item in the theorem from section 4.3, we have

$$X \langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X Z \rangle$$

Permutation of these letters cyclically yields

$$Y \langle Z, X \rangle = \langle \nabla_Y Z, X \rangle + \langle Z, \nabla_Y X \rangle$$

$$Z \langle X, Y \rangle = \langle \nabla_Z X, Y \rangle + \langle X, \nabla_Z Y \rangle$$

Add the first two equations and subtract the third to discover that the expression

$$X \langle Y, Z \rangle + Y \langle Z, X \rangle - Z \langle X, Y \rangle$$

equals

$$\langle \nabla_X Y + \nabla_Y X, Z \rangle + \langle \nabla_X Z - \nabla_Z X, Y \rangle + \langle \nabla_Y Z - \nabla_Z Y, X \rangle$$

Apply the formula $\nabla_X Y - \nabla_Y X = [X, Y]$ and its counterparts using other letters. The last two terms simplify immediately, while the left half of the first term becomes $\nabla_X Y + \nabla_Y X = \nabla_X Y + \{\nabla_X Y - [X, Y]\}$, or $2\nabla_X Y - [X, Y]$. Consequently, the previously displayed formula becomes

$$2 \langle \nabla_X Y, Z \rangle - \langle [X, Y], Z \rangle + \langle [X, Z], Y \rangle + \langle [Y, Z], X \rangle$$

Putting this altogether, we obtain the following important formula

Important Formula:

$$\begin{aligned} 2\langle \nabla_X Y, Z\rangle = & \quad X\langle Y,Z\rangle \quad +Y\langle X,Z\rangle \quad -Z\langle X,Y\rangle \\ & -\langle X,[Y,Z]\rangle \quad -\langle Y,[X,Z]\rangle \quad +\langle Z,[X,Y]\rangle \end{aligned}$$

Continuation of proof: Apply this formula in the special case that $X = \frac{\partial}{\partial u_i}$ and $Y = \frac{\partial}{\partial u_j}$ and $Z = \frac{\partial}{\partial u_k}$. Notice that the Lie bracket of any two such fields is zero. We obtain

$$2\left\langle \sum_l \Gamma^l_{ij}\frac{\partial}{\partial u_l}, \frac{\partial}{\partial u_k}\right\rangle = \frac{\partial}{\partial u_i}\ g_{jk} + \frac{\partial}{\partial u_j}\ g_{ik} - \frac{\partial}{\partial u_k}\ g_{ij}$$

or equivalently

$$\sum_l \Gamma^l_{ij} g_{lk} = \frac{1}{2}\left\{\frac{\partial g_{jk}}{\partial u_i} + \frac{\partial g_{ik}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_k}\right\}$$

and the result clearly follows after multiplying both sides by g^{-1} and summing appropriately. QED.

We have now proved the following extremely important result:

Theorem 40 *Let X and Y be tangent vector fields. The expression $\nabla_X Y$ can be computed intrinsically by a two-dimensional worker living on the surface.*

4.6 Tangent Fields Along Curves

Suppose Y is a tangent field on a surface and $\gamma(t)$ is a curve. We want to compute the derivative of Y *in the direction of the curve.* To do so, define $X = \gamma'(t)$. Then the derivative of the three-dimensional vector field in the direction of the curve is $X(Y)$ and the tangential component of this derivative is $\nabla_X Y$. It is convenient to call the first of these derivatives $\frac{dY}{dt}$ and the second $\frac{DY}{dt}$.

Let us deduce formulas for these derivatives. The formulas will reveal something interesting, namely that the derivatives depend only on Y along the curve, and not on the extension of Y to a global vector field.

We first compute $X(Y)$. Write $Y = (Y_1, Y_2) = Y_1\frac{\partial \vec{s}}{\partial u} + Y_2\frac{\partial \vec{s}}{\partial v}$ as a three dimensional vector $V(u,v)$. Write $\gamma'(t) = \left(\frac{du}{dt}, \frac{dv}{dt}\right)$. Then $X(V) = \frac{du}{dt}\frac{\partial V}{\partial u} + \frac{dv}{dt}\frac{\partial V}{\partial v}$. By the chain rule, this equals $\frac{d}{dt}V(u(t),v(t)) = \frac{d}{dt}V(\gamma(t))$. The answer depends only on V along $\gamma(t)$.

We next compute $\nabla_{\gamma'(t)}Y$. According to our formula, it is

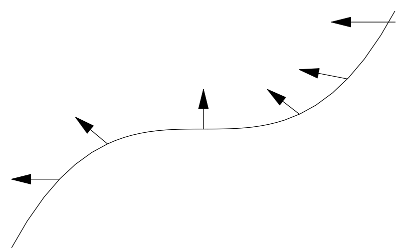
$$\sum_{ik}\frac{d\gamma_i}{dt}\left\{\frac{\partial Y_k}{\partial u_i} + \sum_j \Gamma^k_{ij} Y_j\right\}\frac{\partial}{\partial u_k}$$

By the chain rule, $\frac{d}{dt}Y_k(\gamma(t)) = \frac{\partial Y_k}{\partial u_j}\frac{\partial \gamma_j}{\partial t}$, so the result can be rewritten as

$$\sum_k \left\{ \frac{dY_k}{dt} + \sum_{ij} \Gamma_{ij}^k \frac{d\gamma_i}{dt} Y_j \right\} \frac{\partial}{\partial u_k}$$

This answer also depends only on the values of Y on the curve $\gamma(t)$ and not on the extension of these values to the rest of the surface.

We use this observation to extend our theory to a slightly more general situation. Suppose then that $\gamma(t)$ is a curve and $Y(t)$ is a function which assigns to each t a tangent vector $Y(t)$ starting at $\gamma(t)$. We call Y a *tangent field along the curve.* See the picture below. In coordinates $Y(t) = (Y_1(t), Y_2(t))$. This tangent vector can be converted to a three-dimensional vector $(V_1(t), V_2(t), V_3(t))$.



Definition 20 *The ordinary derivative of $Y(t)$ is a three-dimensional field along the curve, and is defined by*

$$\frac{dY}{dt} = \left(\frac{dV_1}{dt}, \frac{dV_2}{dt}, \frac{dV_3}{dt} \right).$$

The covariant derivative *of $Y(t)$ is a tangent field along the curve, and is defined by*

$$\frac{DY}{dt} = \sum_k \left\{ \frac{dY_k}{dt} + \sum_{ij} \Gamma_{ij}^k \frac{d\gamma_i}{dt} Y_j \right\} \frac{\partial}{\partial u_k}$$

Theorem 41 *Let $\gamma(t)$ be a curve on the surface and let $Y(t)$ be a tangent field along γ.*

1. *The three-dimensional vector $\frac{dY}{dt}$ can be decomposed as follows into a tangential component and a normal component:*

$$\frac{dY}{dt} = \frac{DY}{dt} + b(\gamma'(t), Y(t))\vec{n}$$

2. $\frac{D}{dt}\{r_1 Y(t) + r_2 Z(t)\} = r_1 \frac{DY}{dt} + r_2 \frac{DZ}{dt}$
3. $\frac{d}{dt}\langle Y(t), Z(t)\rangle = \left\langle \frac{DY}{dt}, Z\right\rangle + \left\langle Y, \frac{DZ}{dt}\right\rangle$

Proof: This follows immediately from the corresponding facts for vector fields. QED.

When $\gamma(t)$ is a curve, the derivative $\gamma'(t)$ is a tangent field along the curve. The remaining sections of this chapter are about $\frac{DY}{dt}$ for this special case $Y = \gamma'(t)$.

4.7 Acceleration

Forget our course for a moment and consider the path of a particle moving in R^3. The position of the particle at time t will be $\gamma(t)$ for some curve γ. The velocity of the particle is then $\gamma'(t)$. We want to talk of the speed of the particle at time t, but we cannot call it $s(t)$ because s has a different meaning in this course. We will call the speed $w(t)$. Then $\gamma'(t) = w(t)T(t)$ where $T(t)$ is the unit tangent vector.

Differentiating again, we find that the acceleration is given by

$$\gamma''(t) = \frac{dw}{dt}T(t) + w\frac{dT}{dt}$$

It is convenient to find a more geometrical description of the second term. Write $T(t) = T(s(t))$ as in chapter one, where $T(s)$ is the tangent vector along the curve parameterized by arc length. Then

$$\frac{dT}{dt} = \frac{dT}{ds}\frac{ds}{dt} = (\kappa N)\, w(t).$$

Putting this together with the previous formula, we conclude that

Theorem 42 *If a particle travels along a curve $\gamma(t)$ and has speed $w(t)$ at time t, then its acceleration is given by*

$$\gamma''(t) = \frac{dw}{dt}T + w^2\kappa N$$

where T is the unit tangent vector, N is the unit normal vector, and κ is the curvature of the curve.

Remark: This theorem is well known to anyone who has driven a car. Passengers usually sit calmly. But sometimes they are pressed into the seat or thrown forward in the direction T. The force in this direction is independent of the speed of the car and depends only on acceleration or deceleration (that is, on $\frac{dw}{dt}$).

Drivers are also thrown sidewise. The sidewise force is not caused by accelerating the car, but instead depends on cornering. The magnitude of the sidewise force is proportional to κ, the curvature of the road as determined by the engineer who designed it, and proportional to speed squared. So driving around a corner slowly is a piece of cake, but driving around it fast is dangerous.

4.8 Surface Decomposition of Acceleration

We are going to refine the previous theorem when the car is driven on a surface. In that case, the driver will feel a force in the direction T, caused as before by acceleration or deceleration. The force perpendicular to T can be further decomposed into a component in the direction n and a component in the direction $n \times T$. The component in the direction n presses the car down onto the surface. We will see that it is proportional to speed squared and to the curvature of the surface in the direction T. The component in the direction $n \times T$ pushes the car back and forth across the surface. We will see that it is proportional to the square of the speed and to a number κ_g called the *geodesic curvature* of the curve. This geodesic curvature, which will be computed using the covariant derivative, measures the curvature of the path $\gamma(t)$ from the point of view of a two-dimensional worker in the surface. When the path follows a geodesic, this curvature is zero; otherwise it measures the divergence of the path from a geodesic in the same direction.

Theorem 43 *Let $\gamma(t)$ be a C^∞ curve on a surface. Suppose it has unit tangent vector $T(t)$ and speed $w(t)$.*

1. *The decomposition*

$$\gamma''(t) = \frac{dw}{dt}T + w^2\kappa N$$

   obtained earlier can be refined to a decomposition into three components: one along T, one along $\vec{n}$, and one along $n \times T$ as follows:

$$\vec{\gamma}''(t) = \frac{dw}{dt}\vec{T} + w^2\kappa_g(\vec{n} \times \vec{T}) + w^2 b(T,T)\vec{n}$$

   This expression defines a real-valued quantity κ_g, called the geodesic curvature *of γ.*

2. *We have the following decomposition of γ'' into a component tangent to the surface*

and a component normal to the surface:

$$\gamma''(t) = \frac{D\gamma'}{dt} + w^2 b(T,T)n$$

3. *We have the following decomposition of the surface tangential component of γ'' into a component tangent to the curve and a component normal to the curve:*

$$\frac{D\gamma'}{dt} = \frac{dw}{dt}T + w^2\kappa_g(n \times T).$$

4. *The curvature κ of the curve is thus decomposed into the geodesic curvature κ_g of γ within the surface, and the curvature $b(T,T)$ of the surface itself. Moreover,*

$$\kappa^2 = \kappa_g^2 + b(T,T)^2$$

5. *The normal to the curve $N(t)$ is decomposed into a vector tangent to the surface and a vector normal to the surface. Moreover:*

$$N = \frac{\kappa_g}{\kappa}(n \times T) + \frac{b(T,T)}{\kappa}n.$$

Proof: Let $\gamma(t)$ be an arbitrary C^∞ curve on the surface. Then $\gamma'(t)$ is a tangent field along the curve. The expression $\frac{d}{dt}\gamma'(t)$ defined in section 4.6 is just the acceleration $\gamma''(t)$. According to results in that section, the tangential component of this acceleration vector is

$$\frac{D\gamma'(t)}{dt} = \sum_k \left\{ \frac{d^2\gamma_k(t)}{dt^2} + \sum_{ij} \Gamma_{ij}^k \frac{d\gamma_i}{dt}\frac{d\gamma_j}{dt} \right\} \frac{\partial}{\partial u_k}$$

and the normal component is

$$b(\gamma'(t), \gamma'(t))\vec{n}$$

The tangential component $\frac{D\gamma'}{dt}$ can be decomposed into a component in the direction T and a component in the direction $n \times T$. To obtain this decomposition, begin by differentiating both sides of the equation $w^2 = \langle \gamma'(t), \gamma'(t) \rangle$ with respect to t, to obtain

$$2w\frac{dw}{dt} = \frac{d}{dt}\left\langle \gamma'(t), \gamma'(t) \right\rangle = \left\langle \frac{D\gamma'}{dt}, \gamma'(t) \right\rangle + \left\langle \gamma'(t), \frac{D\gamma'}{dt} \right\rangle = 2\left\langle \frac{D\gamma'}{dt}, \gamma'(t) \right\rangle$$

or $2w\frac{dw}{dt} = 2\left\langle \frac{D\gamma'}{dt}, wT \right\rangle$, and so

$$\frac{dw}{dt} = \left\langle \frac{D\gamma'}{dt}, T \right\rangle$$

Consequently, the tangential component of this decomposition is $\frac{dw}{dt}T$. Comparing the two decompositions in the first part of the theorem gives

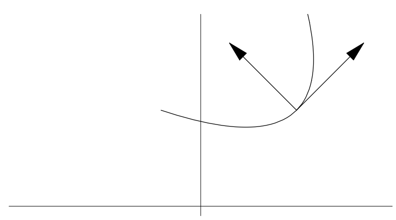
$$\kappa N = \kappa_g(n \times T) + b(T,T)n$$

and the rest of the theorem easily follows. QED.

4.9 Geodesic Curvature

In the previous section, we defined an important quantity called *geodesic curvature.* It is the curvature of $\gamma(t)$ within the surface, as computed by a two-dimensional worker who believes that geodesics are straight. The quantity κ_g can be positive or negative.

Geodesic curvature depends on an orientation of the surface, since it depends on the vector $n \times T$. If we are in the plane, we always suppose that n points upward in the positive z direction. In this case, T and $n \times T$ form a right handed coordinate system in the sense that we move counterclockwise to get from T to $n \times T$. We will always adopt this convention. We look down on the surface from the tip of the n vector and then T must be rotated counterclockwise by ninety degrees to reach $n \times T$. The picture below shows T and $n \times T$ and a curve with positive κ_g.



However, it is important to remember that T and $n \times T$ are perpendicular unit vectors *on the surface*, and not necessarily in local coordinates, where the g_{ij} distort distances and angles.

Theorem 44 *Let $\gamma(t)$ be a curve on the surface.*

1. *Geodesic curvature can be computed extrinsically using the formula*

$$\kappa_g = \frac{\frac{d^2\gamma}{dt^2} \cdot \left(n \times \frac{d\gamma}{dt}\right)}{\frac{d\gamma}{dt} \cdot \frac{d\gamma}{dt}}$$

2. *Geodesic curvature can also be computed intrinsically. The expression*

$$\frac{\frac{D\gamma'}{dt} - \left\langle \frac{D\gamma'}{dt}, T \right\rangle T}{\langle \gamma', \gamma' \rangle}$$

equals $\kappa_g(n \times T)$. If the displayed vector is obtained from T by rotating counterclockwise, then κ_g is the length of this vector. Otherwise, γ_g is the negative of this length.

3. *In the particular case when $\gamma(t)$ is parameterized proportional to arc length, the previous formula simplies to*

$$\frac{\frac{D\gamma'}{dt}}{\langle \gamma', \gamma' \rangle} = \frac{\sum_k \left\{ \frac{d^2\gamma_k}{dt^2} + \sum_{ij} \Gamma_{ij}^k \frac{d\gamma_i}{dt} \frac{d\gamma_j}{dt} \right\} \frac{\partial}{\partial u_k}}{\sum_{ij} g_{ij} \frac{d\gamma_i}{dt} \frac{d\gamma_j}{dt}}$$

4. *In particular, a curve parameterized proportional to arc length has geodesic curvature zero exactly when it is a geodesic.*

Proof: This theorem immediately follows from earlier results in the chapter.

4.10 Example

Consider the sphere parameterized by spherical coordinates: $s(\theta, \phi) = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)$. Then

$$\frac{\partial s}{\partial \theta} = (-\sin\phi\sin\theta, \sin\phi\cos\theta, 0)$$

$$\frac{\partial s}{\partial \phi} = (\cos\phi\cos\theta, \cos\phi\sin\theta, -\sin\phi)$$

$$g_{11} = \sin^2\phi \qquad g_{12} = 0 \qquad g_{22} = 1.$$

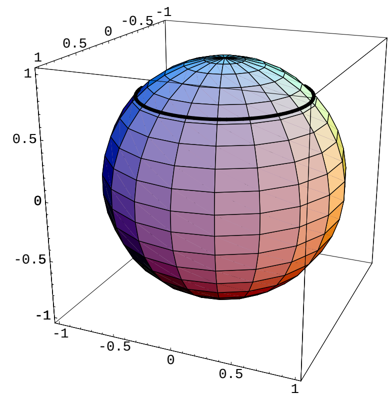
Consequently,

$$\frac{\partial s}{\partial \theta} \times \frac{\partial s}{\partial \phi} = -\sin\phi(\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi),$$

which points inward. We prefer to orient the sphere with outward pointing normal, so we will change the sign of n. After normalizing we have

$$n = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi).$$

Consider the curves on the sphere of constant latitude, $\gamma(t) = (\sin\phi\cos t, \sin\phi\sin t, \cos\phi)$. Let us compute the geodesic curvature of these curves extrinsically.



We have $\gamma' = (-\sin\phi\sin t, \sin\phi\cos t, 0)$ and $\gamma'' = (-\sin\phi\cos t, -\sin\phi\sin t, 0)$. So $\gamma'\cdot\gamma' = \sin^2\phi$. Along the curve, $n = (\sin\phi\cos t, \sin\phi\sin t, \cos\phi)$ and so

$$n \times \gamma' = (-\sin\phi\cos\phi\cos t, -\sin\phi\cos\phi\sin t, \sin^2\phi).$$

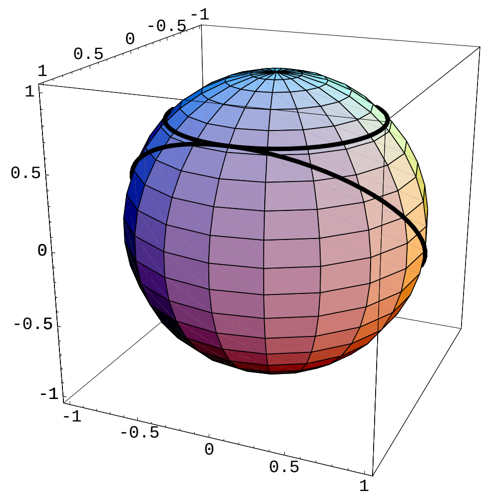
We want to compute $n \times T$ along the curve, so we divide by the length of γ', that is, $\sin\phi$:

$$n \times T = (-\cos\phi\cos t, -\cos\phi\sin t, \sin\phi)$$

Thus

$$\kappa_g = \frac{\frac{d^2\gamma}{dt^2}\cdot\left(n\times\frac{d\gamma}{dt}\right)}{\frac{d\gamma}{dt}\cdot\frac{d\gamma}{dt}} = \frac{\cos\phi}{\sin\phi}.$$

The picture below shows that our formula $\kappa_g = \frac{\cos\phi}{\sin\phi}$ is reasonable. According to the result, $\kappa_g = 0$ on the equator when $\phi = \pi/2$ and $\cos\phi = 0$. We expect this because the equator is a geodesic. On the other hand, the geodesic curvature becomes positive as we consider latitude curves higher on the sphere, since these paths curve away in the positive direction from the corresponding geodesic on the sphere. At the very top of the picture, the latitude curve is a very small circle and curves away from the geodesic at an extremely rapid rate, so $\kappa_g = \frac{\cos\phi}{\sin\phi}$ approaches infinity.



Let us repeat this calculation intrinsically. All Γ^k_{ij} vanish except

$$\Gamma^1_{12} = \frac{\cos\phi}{\sin\phi} \qquad \Gamma^2_{11} = -\sin\phi\cos\phi$$

In local coordinates, our curve is $\gamma(t) = (t, \phi)$ and $\gamma'(t) = (1, 0)$. The first component of $\frac{D\gamma'}{dt}$ is

$$\frac{d^2\gamma_1}{dt^2} + 2\Gamma^1_{12}\frac{d\gamma_1}{dt}\frac{d\gamma_2}{dt} = \frac{d}{dt}(1) + \frac{2\cos\phi}{\sin\phi}0 = 0.$$

The second component of $\frac{D\gamma'}{dt}$ is

$$\frac{d^2\gamma_2}{dt^2} + \Gamma^2_{11}\frac{d\gamma_1}{dt}\frac{d\gamma_1}{dt} = \frac{d}{dt}(0) - \sin\phi\cos\phi(1)^2 = -\sin\phi\cos\phi.$$

The expression $\langle\gamma', \gamma'\rangle$ equals

$$g_{11}1^2 = \sin^2\phi.$$

Since γ moves at constant speed, we must calculate

$$\frac{\frac{D\gamma'}{dt}}{\langle\gamma',\gamma'\rangle} = \frac{\sum_k \left\{ \frac{d^2\gamma_k}{dt^2} + \sum_{ij} \Gamma^k_{ij} \frac{d\gamma_i}{dt}\frac{d\gamma_j}{dt} \right\} \frac{\partial}{\partial u_k}}{\sum_{ij} g_{ij} \frac{d\gamma_i}{dt}\frac{d\gamma_j}{dt}} = \frac{(0, -\sin\phi\cos\phi)}{\sin^2\phi}$$

This vector equals $\left(0, -\frac{\cos\phi}{\sin\phi}\right)$. Its length is thus the geodesic curvature up to a sign. The length of this vector squared is

$$g_{11}\cdot 0 + 2g_{12}\cdot 0 + g_{22}\frac{\cos^2\phi}{\sin^2\phi} = \frac{\cos^2\phi}{\sin^2\phi}$$

and so the geodesic curvature is

$$\kappa_g = \pm\frac{\cos\phi}{\sin\phi}.$$

We must determine the correct sign. In the two-dimensional coordinate space (θ,ϕ), our curve has tangent $(1,0)$ and the vector computed above, $\left(0, -\frac{\cos\phi}{\sin\phi}\right)$, points downward. Hence we get to it by rotating clockwise. But we are supposed to get to $n \times T$ by rotating counterclockwise, so we should choose the minus sign.

But wait. At the start of this section, we oriented the sphere using outward pointing normals, warning that the expression $\frac{\partial s}{\partial\theta} \times \frac{\partial s}{\partial\phi}$ points in the wrong direction. So the orientation of our coordinate space is opposite the orientation chosen on the sphere, and we must compensate for this change. Consequently, we should choose the plus sign.

Chapter 5

The Theorema Egregium

5.1 Introduction

We have divided the theory of surfaces into two pieces. The intrinsic piece can be understood by a two-dimensional worker on the surface, and is determined by the first fundamental form g_{ij}. Intrinsic theory is very rich, allowing two dimensional workers to find geodesics, to compute derivatives of vector fields $\nabla_X Y$, to determine acceleration $\frac{D\gamma'}{dt}$, and to compute geodesic curvature κ_g.

The extrinsic piece of surface theory describes the curvature of the surface into the third dimension. This curvature is determined by the second fundamental form b_{ij}.

Let us compare this theory with the theory of curves in chapter one. Suppose a one-dimensional worker lives on a curve. In one-dimension, there is only one possible geometry, the geometry of the real line. So instead of constructing an elaborate intrinsic theory, we parameterized the curve by arc length to directly reflect this geometry.

The numbers κ and τ are the curve theoretic analogues of the b_{ij} in surface theory.

According to the fundamental theorem of curve theory, a curve is completely determined by κ and τ up to Euclidean motion. This theorem was proved using the existence theorem for ordinary differential equations.

There is an analogous theorem for surfaces, although we probably will not prove it. The theorem asserts that the g_{ij} and b_{ij} completely determine the surface up to Euclidean motion. This time, the theorem is proved using the existence theorem for partial differential equations.

The theory of ordinary differential equations and the theory of partial differential equations differ in an important respect. In ordinary theory, all reasonable equations have solutions. But most partial differential equations do not have solutions unless they satisfy an extra *integrability condition.* For example, the simplest partial differential equation is the following system for an unknown $f(x, y)$:

$$\begin{cases} \frac{\partial f}{\partial x} & = & E_x(x, y) \\ \frac{\partial f}{\partial y} & = & E_y(x, y) \end{cases}$$

and this system has solutions exactly when

$$\frac{\partial E_y}{\partial x} = \frac{\partial E_x}{\partial y}$$

This chapter is about the integrability conditions in surface theory — conditions relating the g_{ij} and b_{ij} which must hold before a surface exists with these fundamental forms. We will determine all of these conditions (there are two). One of them will give us Gauss's theorema egregium.

5.2 The Fundamental Theorem

In ordinary calculus, the partials $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ were introduced. An extremely important theorem asserts that these operators commute, so

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.$$

On surfaces, there are no natural directions and we are forced to replace the two partials with vector fields X and Y. These vector fields usually have non-constant coefficients, so the operators X and Y do not commute. Instead, the non-commutativity $X(Y(f)) - Y(X(f))$ is measured by another vector field $[X, Y]$. We have seen this field appear several times:

$$\begin{aligned} X(Y(f)) - Y(X(f)) &= [X, Y]f \\ X(Y) - Y(X) &= [X, Y] \\ \nabla_X Y - \nabla_Y X &= [X, Y] \end{aligned}$$

You might expect one other formula:

$$\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z = \nabla_{[X,Y]} Z$$

However, this formula is not quite correct. The correct formula will lead us to Gauss's great theorem.

Theorem 45 *Let X and Y be tangent vector fields on a surface, and let V by an arbitrary three-dimensional field on the surface. Then*

$$X(Y(V)) - Y(X(V)) = [X, Y](V)$$

Proof: Write $V = (V_1, V_2, V_3)$. The definition of $X(V)$ states that we should differentiate each coordinate separately:

$$X(V) = (\ X(V_1),\ X(V_2),\ X(V_3)\ )$$

But then our theorem is just the assertion that for functions f we have $X(Y(f)) - Y(X(f)) = [X, Y]f$. QED.

Remark: We reach the inner sanctum of surface theory by decomposing this equation into normal and tangential components. The resulting equations are given next. You may wish to deduce the equations yourself without looking.

Theorem 46 (Fundamental Equations of Surface Theory) *Let $X, Y, Z,$ and W be tangent fields on a surface. The following relations hold between $\nabla, b,$ and B:*

1. $\left\langle \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z,\ W \right\rangle = -\left\{ b(X, Z) b(Y, W) - b(Y, Z) b(X, W) \right\}$
2. $b(X, \nabla_Y Z) + X(b(Y, Z)) - b(Y, \nabla_X Z) - Y(b(X, Z)) = b([X, Y], Z)$
3. $\nabla_X B(Y) - \nabla_Y B(X) = B([X, Y])$

Proof: We know that $X(Y(Z)) - Y(X(Z)) = [X, Y]Z$. Let us compute the tangential component of this equation. We have

$$X(Y(Z)) = X(\nabla_Y Z + b(Y, Z)n).$$

But

$$X(\nabla_Y Z) = \nabla_X \nabla_Y Z + b(X, \nabla_Y Z)n$$

and

$$X(\ b(Y, Z)n\ ) = X(b(Y, Z))n + b(Y, Z)X(n) = X(b(Y, Z))n + b(Y, Z)B(X).$$

So we have

$$X(Y(Z)) = \left\{ \nabla_X \nabla_Y Z + b(Y, Z)B(X) \right\} + \left\{ b(X, \nabla_Y Z) + X(b(Y, Z)) \right\} n.$$

Subtract the same expression with X and Y interchanged. The tangential component of the result is

$$\nabla_X \nabla_Y Z + b(Y, Z)B(X) - \nabla_Y \nabla_X Z - b(X, Z)B(Y)$$

and must equal the tangential component of $[X, Y]Z$, which is $\nabla_{[X,Y]} Z$. Therefore

$$\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z = b(X, Z)B(Y) - b(Y, Z)B(X).$$

Take the inner product of this equation with W to obtain the first part of the theorem.

The second part of the theorem follows similarly by equating normal components of $X(Y(Z)) - Y(X(Z)) - [X, Y]Z$.

To obtain the third part of the theorem, decompose the equation $X(Y(n)) - Y(X(n)) = [X, Y](n)$ into tangential and normal components. Notice that

$$X(Y(n)) = X(B(Y)) = \nabla_X B(Y) + n(X, B(Y))n.$$

If we subtract the same equation with X and Y interchanged, we get

$$X(Y(n)) - Y(X(n)) = \Big\{\nabla_X B(Y) - \nabla_Y B(X)\Big\} + \Big\{n(X, B(Y)) - n(Y, B(X))\Big\}$$

and the result should equal $[X, Y](n) = B([X, Y])$. The normal component of this equation is zero by the symmetry of the operator B, and the tangential component gives the third result of the theorem. QED.

Remark: The above equations are formal and algebraic. It took genius to realize that they hide important geometrical facts. We'll reveal these facts in the next sections.

Incidentally, our convention that X, Y, and Z are tangent vectors and U, V, and W are arbitrary three-dimensional vectors will not work in this section and the sections which follow because we need four tangent vectors. From now on, W is a tangent vector!

5.3 Gaussian Curvature

In this section, we will study the right side of the first equation of the previous theorem and try to disentangle the algebra. It will turn out that the information in the right side is essentially the single number $\kappa_1 \kappa_2$.

It is temporarily convenient to introduce the notation

$$b(X, Y, Z, W) = b(X, Z)b(Y, W) - b(Y, Z)b(X, W).$$

Notice that the expression $b(X,Y,Z,W)$ is linear in each variable separately if we hold the others fixed. So if we define

$$b_{ijkl} = b\left(\frac{\partial}{\partial u_i}, \frac{\partial}{\partial u_j}, \frac{\partial}{\partial u_k}, \frac{\partial}{\partial u_l}\right)$$

then

$$b(X,Y,Z,W) = \sum_{ijkl} b_{ijkl} X_i Y_j Z_k W_l$$

Notice also that $b(X,Y,Z,W)$ changes sign if we interchange X and Y, or if we interchange Z and W. In particular, the expression is zero if $X = Y$ or if $Z = W$. So the only nonzero coefficients b_{ijkl} are

$$b_{1212} = -b_{2112} = b_{2121} = -b_{1221}$$

and

$$B(X,Y,Z,W) = b_{1212}\Big(X_1Y_2Z_1W_2 - X_2Y_1Z_1W_2 + X_2Y_1Z_2W_1 - X_1Y_2Z_2W_1\Big).$$

Therefore $b(X,Y,Z,W)$ depends on a single number. It is convenient to get our hands on this coefficient in a more invariant manner.

Theorem 47 *Fix a point p on the surface.*

1. *Let X and Y be orthonormal vectors at p. Then the number*

$$b(X,X)b(Y,Y) - b(X,Y)^2$$

is independent of the choice of orthonormal basis.

2. *The number equals $\kappa_1\kappa_2$.*

3. *The coefficient b_{1212} introduced earlier equals*

$$\det(g_{ij})\ \kappa_1\kappa_2.$$

Proof: These facts follow from earlier results stating that $\det B = \kappa_1\kappa_2$ and $B = -g^{-1}b$. We'll give a direct proof for completeness. Choose e_1 and e_2 principal directions. Then $B(e_1) = \kappa_1 e_1$ and $B(e_2) = \kappa_2 e_2$. So $b(e_1,e_1) = \langle B(e_1), e_1\rangle = \kappa_1 \langle e_1, e_1\rangle = \kappa_1$, etc. Write $X = a_{11}e_1 + a_{21}e_2$ and $Y = a_{12}e_1 + a_{22}e_2$. Then by linearity

$$b(X,X) = a_{11}^2 b(e_1,e_1) + 2a_{11}a_{21}b(e_1,e_2) + a_{21}^2 b(e_2,e_2) = a_{11}^2\kappa_1 + a_{21}^2\kappa_2.$$

Similarly $b(X,Y) = a_{11}a_{12}\kappa_1 + a_{21}a_{22}\kappa_2$ and $b(Y,Y) = a_{12}^2\kappa_1 + a_{22}^2\kappa_2$. A short calculation then gives

$$b(X,X)b(Y,Y) - b(X,Y)^2 = (a_{11}a_{22} - a_{12}a_{21})^2\ \kappa_1\kappa_2.$$

However,

$$\langle X, X\rangle = a_{11}^2 \langle e_1, e_1\rangle + 2a_{11}a_{21} \langle e_1, e_2\rangle + a_{21}^2 \langle e_2, e_2\rangle = a_{11}^2 + a_{21}^2.$$

Similarly $\langle X, Y\rangle = a_{11}a_{12} + a_{21}a_{22}$ and $\langle Y, Y\rangle = a_{12}^2 + a_{22}^2$, and another short calculation gives

$$\langle X, X\rangle \langle Y, Y\rangle - \langle X, Y\rangle^2 = (a_{11}a_{22} - a_{12}a_{21})^2 .$$

Consequently

$$\kappa_1\kappa_2 = \frac{b(X, X)b(Y, Y) - b(X, Y)^2}{\langle X, X\rangle \langle Y, Y\rangle - \langle X, Y\rangle^2}$$

If X and Y are orthonormal, the denominator is one and we have

$$\kappa_1\kappa_2 = b(X, X)b(Y, Y) - b(X, Y)^2.$$

If $X = \frac{\partial}{\partial u}$ and $Y = \frac{\partial}{\partial v}$ then the denominator is $\left(g_{11}g_{22} - g_{12}^2\right)$ and

$$b_{1212} = b\left(\frac{\partial}{\partial u}, \frac{\partial}{\partial v}, \frac{\partial}{\partial u}, \frac{\partial}{\partial v}\right) = b(X, Y, X, Y),$$

which equals

$$b(X, X)b(Y, Y) - b(X, Y)^2 = \left(g_{11}g_{22} - g_{12}^2\right) \kappa_1\kappa_2.$$

The theorem follows. QED.

Theorem 48 (The Theorema Egregium) *The number $\kappa_1\kappa_2$ can be computed intrinsically by a two-dimensional worker living on the surface.*

Proof: Choose an orthonormal basis X, Y of tangent vectors at p. Extend X and Y to vector fields on the surface. By the first equation of surface theory,

$$\kappa_1\kappa_2 = -\left\langle \nabla_X \nabla_Y X - \nabla_Y \nabla_X X - \nabla_{[X,Y]} X,\ Y \right\rangle$$

QED.

5.4 The Curvature Tensor

It is useful to give a systematic treatment of the intrinsic approach to $\kappa_1\kappa_2$. Everything we say in this section works in higher dimensions, and in situations when the object is not embedded in Euclidean space.

Definition 21 *Let X, Y, Z, W be tangent vectors at a point p on a surface. Extend these vectors to vector fields on the surface. Define*

$$R(X, Y, Z, W) = \left\langle \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z,\ W \right\rangle$$

Then the map R, which assigns a number to any combination of four tangent vectors, is called the Riemann curvature tensor.

Theorem 49 *Let X, Y, Z, and W be tangent vectors at p.*

1. *The number $R(X, Y, Z, W)$ depends only on the values of X, Y, Z, and W at p, and not on their extension to vector fields.*
2. *The map $R(X, Y, Z, W)$ is linear in each variable separately if the others are held fixed.*
3. *There are numbers R_{ijkl} so that*

$$R(X, Y, Z, W) = \sum_{ijkl} R_{ijkl} X_i Y_j Z_k W_l$$

4. *The R_{ijkl} are given by*

$$R_{ijkl} = \sum_s \left\{ \frac{\partial \Gamma^s_{jk}}{\partial u_i} - \frac{\partial \Gamma^s_{ik}}{\partial u_j} + \sum_m \Gamma^m_{jk} \Gamma^s_{im} - \sum_m \Gamma^m_{ik} \Gamma^s_{jm} \right\} g_{sl}$$

5. *The number $R(X, Y, Z, W)$ changes sign if X and Y are interchanged, or if Z and W are interchanged.*
6. *The only non-zero coefficients R_{ijkl} are*

$$R_{1212} = -R_{2112} = R_{2121} = -R_{1221}.$$

7. *Thus $R(X, Y, Z, W)$ contains only one piece of information. This information is*

$$R_{1212} = -\det(g_{ij})\ \kappa_1 \kappa_2.$$

8. *Let e_1, e_2 be an orthonormal basis of tangent vectors at p. The number $R(e_1, e_2, e_1, e_2)$ does not depend on the choice of this basis.*
9. *$R(e_1, e_2, e_1, e_2) = -\kappa_1 \kappa_2$.*

Proof: Let us start with vector fields X, Y, Z, and W. The expression $R(X, Y, Z, W)$ is linear over the reals in each variable separately. We are going to prove that $R(fX, Y, Z, W) = fR(X, Y, Z, W)$ for any C^∞ function f, and that a similar result holds if we multiple any of Y, Z, or W by f. It immediately follows that

$$R(X, Y, Z, W) = \sum_{ijkl} R\left(\frac{\partial}{\partial u_i}, \frac{\partial}{\partial u_j}, \frac{\partial}{\partial u_k}, \frac{\partial}{\partial u_l}\right) X_i Y_j Z_k W_l,$$

and consequently that $R(X, Y, Z, W)$ at a point p depends only on the values of X, Y, Z, and W at p, since no derivatives of the coefficient functions appear in the final expression.

As an initial step, notice that for any C^∞ function g we have

$$[fX, Y](g) = fX(Y(g)) - Y(fX(g)) = fX(Y(g)) - Y(f)X(g) - fY(X(g))$$

This expression equals $\Big\{f[X, Y] - Y(f)X\Big\}g$. Since the equation holds for all functions g, we have

$$[fX, Y] = f[X, Y] - Y(f)X.$$

Now use the product rule for the covariant derivative, proved as point 2 of the first theorem in section 4.4, to conclude that

$$\nabla_{fX} \nabla_Y Z - \nabla_Y \nabla_{fX} Z - \nabla_{[fX,Y]} Z$$

equals

$$f\nabla_X \nabla_Y Z - Y(f)\nabla_X Z - f\nabla_Y \nabla_X Z - \nabla_{f[X,Y]-Y(f)X} Z$$

which in turn equals

$$f\nabla_X \nabla_Y Z - Y(f)\nabla_X Z - f\nabla_Y \nabla_X Z - f\nabla_{[X,Y]} Z + Y(f)\nabla_X Z.$$

This expression simplifies to

$$f\left\{\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z\right\}.$$

So the f has factored out and $R(fX, Y, Z, W) = f\ R(X, Y, Z, W)$.

Clearly $R(X, Y, Z, W) = -R(Y, X, Z, W)$, so it follows that functions f also factor out of the second position. It is immediately clear that functions factor out of the final position because W is only dotted at the very end and certainly never differentiated. To show that f can be factored from the third position, we'll prove that $R(X, Y, Z, W) = -R(X, Y, W, Z)$.

Recall that $Y\langle Z, W\rangle = \langle \nabla_Y Z, W\rangle + \langle Z, \nabla_Y W\rangle$. Differentiating again, we obtain

$$X(Y(\langle Z, W\rangle)) = \langle \nabla_X \nabla_Y Z, W\rangle + \langle \nabla_Y Z, \nabla_X W\rangle + \langle \nabla_X Z, \nabla_Y W\rangle + \langle Z, \nabla_X \nabla_Y W\rangle$$

Subtract the corresponding equation with X and Y interchanged to discover that

$$X(Y(\langle Z, W\rangle)) - Y(X(\langle Z, W\rangle))$$

is equal to

$$\langle \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z, W\rangle + \langle Z, \nabla_X \nabla_Y W - \nabla_Y \nabla_X W\rangle$$

Add $[X, Y]\langle Z, W\rangle$ to the pot to discover that

$$X(Y(\langle Z, W\rangle)) - Y(X(\langle Z, W\rangle)) - [X, Y]\langle Z, W\rangle$$

is equal to

$$\left\langle \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z, W\right\rangle + \left\langle Z, \nabla_X \nabla_Y W - \nabla_Y \nabla_X W - \nabla_{[X,Y]} W\right\rangle$$

However, the top equation is zero because $g = \langle Z, W\rangle$ is a function and for any function whatever, $X(Y(g)) - Y(X(g)) - [X, Y](g) = 0$. So the bottom equation, which equals $R(X, Y, Z, W) + R(X, Y, W, Z)$, is zero.

The remaining parts of the theorem now follow easily. We'll prove only the formula for R_{ijkl}. Notice that the Lie bracket $\left[\frac{\partial}{\partial u_i}, \frac{\partial}{\partial u_j}\right]$ is zero. So

$$\nabla_{\frac{\partial}{\partial u_i}} \nabla_{\frac{\partial}{\partial u_j}} \frac{\partial}{\partial u_k} - \nabla_{\frac{\partial}{\partial u_j}} \nabla_{\frac{\partial}{\partial u_i}} \frac{\partial}{\partial u_k} - \nabla_{\left[\frac{\partial}{\partial u_i}, \frac{\partial}{\partial u_j}\right]} \frac{\partial}{\partial u_k}$$

equals

$$\nabla_{\frac{\partial}{\partial u_i}} \left\{ \sum_s \Gamma_{jk}^s \frac{\partial}{\partial u_s} \right\} - \nabla_{\frac{\partial}{\partial u_j}} \left\{ \sum_s \Gamma_{ik}^s \frac{\partial}{\partial u_s} \right\}$$

which equals

$$\sum_s \left\{ \frac{\partial \Gamma_{jk}^s}{\partial u_i} - \frac{\partial \Gamma_{ik}^s}{\partial u_j} + \sum_m \Gamma_{jk}^m \Gamma_{im}^s - \sum_m \Gamma_{ik}^m \Gamma_{jm}^s \right\} \frac{\partial}{\partial u_s}.$$

The theorem follows by dotting this final result with $\frac{\partial}{\partial u_l}$.

5.5 The Curvature of the Poincare Disk

The results of the previous sections allow us to compute the Gaussian curvature of the Poincare disk. Recall that we defined g_{ij} directly on this disk, without embedding the disk as a surface in R^3. So we cannot calculate $\kappa_1 \kappa_2$ extrinsically.

We parameterize the disk using polar coordinates $s(r,\theta) = (r\cos\theta, r\sin\theta)$. Earlier we gave the metric in rectangular coordinates. These convert to polar coordinates as follows:

$$ds^2 = \frac{4\,\left(dx^2+dy^2\right)}{\left(1-(x^2+y^2)\right)^2} = \frac{4\,\left(dr^2+(rd\theta)^2\right)}{\left(1-r^2\right)^2}$$

Consequently

$$g_{11} = \frac{4}{\left(1-r^2\right)^2} \qquad g_{12} = 0 \qquad g_{22} = \frac{4r^2}{\left(1-r^2\right)^2}$$

A brief calculation shows that the only nonzero Christoffel symbols are

$$\Gamma_{11}^1 = \frac{2r}{1-r^2} \qquad \Gamma_{22}^1 = -r - \frac{2r^3}{1-r^2} \qquad \Gamma_{12}^2 = \frac{1}{r} + \frac{2r}{1-r^2}$$

We must compute $R_{1212} = \sum_s \{\text{fancy expression with } s\}\, g_{s2}$. The only term that matters occurs when $s = 2$ and we obtain

$$R_{1212} = \left\{ \frac{\partial \Gamma_{21}^2}{\partial u_1} - \frac{\partial \Gamma_{11}^2}{\partial u_2} + \sum_m \Gamma_{21}^m \Gamma_{1m}^2 - \sum_m \Gamma_{11}^m \Gamma_{2m}^2 \right\} g_{22}$$

In this expression, $u_1 = r$ and $u_2 = \theta$, so the expression simplifies to

$$R_{1212} = \left\{ \frac{\partial \Gamma_{12}^2}{\partial r} + \Gamma_{12}^2 \Gamma_{12}^2 - \Gamma_{11}^1 \Gamma_{12}^2 \right\} g_{22}$$

The Gaussian curvature is then

$$-\frac{R_{1212}}{\det(g_{ij})} = -\frac{\left(1-r^2\right)^2}{4} \left\{ \frac{\partial \Gamma_{12}^2}{\partial r} + \Gamma_{12}^2 \Gamma_{12}^2 - \Gamma_{11}^1 \Gamma_{12}^2 \right\}$$

When written out, this becomes

$$-\frac{\left(1-r^2\right)^2}{4} \left\{ \frac{\partial}{\partial r}\left(\frac{1}{r} + \frac{2r}{1-r^2}\right) + \left(\frac{1}{r} + \frac{2r}{1-r^2}\right)^2 - \frac{2r}{1-r^2}\left(\frac{1}{r} + \frac{2r}{1-r^2}\right) \right\}$$

which equals

$$-\frac{\left(1-r^2\right)^2}{4} \left\{ \frac{-1}{r^2} + \frac{2+2r^2}{\left(1-r^2\right)^2} + \frac{1}{r^2} + \frac{4}{1-r^2} + \frac{4r^2}{\left(1-r^2\right)^2} - \frac{2}{1-r^2} - \frac{4r^2}{\left(1-r^2\right)^2} \right\}$$

or

$$-\frac{\left(1-r^2\right)^2}{4} \left\{ \frac{2+2r^2}{\left(1-r^2\right)^2} + \frac{2}{1-r^2} \right\}$$

or

$$-\frac{1}{4}\left\{2+2r^2+2-2r^2\right\} = -1.$$

We have proved

Theorem 50 *The Gaussian curvature of the Poincare disk is* -1.

5.6 The Codazzi-Mainardi Equations

This optional section will not be used in the future. We'd like to simplify the remaining equations of surface theory. Recall that these equation are

$$b(X, \nabla_Y Z) + X(b(Y, Z)) - b(Y, \nabla_X Z) - Y(b(X, Z)) = b([X, Y], Z)$$

$$\nabla_X B(Y) - \nabla_Y B(X) = B([X, Y])$$

Theorem 51 *The second equation implies the first.*

Proof: We have

$$X(b(Y, Z)) = -X \langle B(Y), Z \rangle = -\langle \nabla_X B(Y), Z \rangle - \langle B(Y), \nabla_X Z \rangle .$$

Rewriting the last term and moving it to the left gives

$$X(b(Y, Z)) - b(Y, \nabla_X Z) = \langle \nabla_X B(Y), Z \rangle$$

The left side of the first surface equation contains this term and the negative of the term with X and Y interchanged, so the first surface equation can be rewritten

$$\langle \nabla_X B(Y), Z \rangle - \langle \nabla_Y B(Y), Z \rangle = b([X, Y], Z).$$

This is true for all Z exactly when

$$\nabla_X B(Y) - \nabla_Y B(X) = b([X, Y]).$$

QED.

Definition 22 *Let X and Y be vector fields. Then $S(X, Y)$ is the new vector field defined by*

$$S(X, Y) = \nabla_X B(Y) - \nabla_Y B(X) - b([X, Y])$$

Theorem 52 *Let X and Y be tangent vectors at p. Extend X and Y to tangent vector fields.*

1. *The value of $S(X, Y)$ at p depends only on X and Y at p, and not on their extension to vector fields.*
2. *$S(X, Y) = -S(Y, X)$.*
3. *There is a vector $S_{12} = S\left(\frac{\partial}{\partial u}, \frac{\partial}{\partial v}\right)$ such that*

$$S(X, Y) = S_{12}(X_1 Y_2 - X_2 Y_1).$$

4. *We have*

$$S_{12} = \nabla_{\frac{\partial}{\partial u}} B\left(\frac{\partial}{\partial v}\right) - \nabla_{\frac{\partial}{\partial v}} B\left(\frac{\partial}{\partial u}\right)$$

5. *The second fundamental equation of surface theory is the equation $S_{12} = 0$.*

Proof: This entire theorem follows from our standard techniques once we prove that $S(fX, Y) = f\ S(X, Y)$ and $S(X, fY) = f\ S(X, Y)$ for C^∞ functions f. Since we clearly have $S(X, Y) = -S(Y, X)$, it suffices to prove the first of these assertions. But

$$S(fX, Y) = \nabla_{fX} B(Y) - \nabla_Y B(fX) - b([fX, Y]).$$

Using formulas established in the section on the curvature tensor, this becomes

$$f\nabla_X B(Y) - Y(f)B(X) - f\nabla_Y B(X) - b(f[X, Y] - Y(f)X)$$

which simplifies to

$$f\nabla_X B(Y) - Y(f)B(X) - f\nabla_Y B(X) - f\ b([X, Y]) + Y(f)B(X) = f\ S(X, Y).$$

QED.

Remark: We are interested in the equation $S_{12} = 0$. This equation can still be simplified a little more.

Theorem 53 *The vector S_{12} is zero if and only if the equation below is true for $Z = \frac{\partial}{\partial u}$ and for $Z = \frac{\partial}{\partial v}$:*

$$\frac{\partial}{\partial u}\ b\left(\frac{\partial}{\partial v}, Z\right) - b\left(\frac{\partial}{\partial v}, \nabla_{\frac{\partial}{\partial u}} Z\right) = \frac{\partial}{\partial v}\ b\left(\frac{\partial}{\partial u}, Z\right) - b\left(\frac{\partial}{\partial u}, \nabla_{\frac{\partial}{\partial v}} Z\right)$$

Proof: We claim that these equations are equivalent to the statement $\langle S_{12}, Z\rangle = 0$ for these two Z. To prove the assertion, notice that

$$\langle S_{12}, Z\rangle = \left\langle \nabla_{\frac{\partial}{\partial u}} B\left(\frac{\partial}{\partial v}\right) - \nabla_{\frac{\partial}{\partial v}} B\left(\frac{\partial}{\partial u}\right), Z\right\rangle$$

But

$$\frac{\partial}{\partial u}\ b\left(\frac{\partial}{\partial v}, Z\right) = -\frac{\partial}{\partial u}\left\langle B\left(\frac{\partial}{\partial v}\right), Z\right\rangle = -\left\langle \nabla_{\frac{\partial}{\partial u}} B\left(\frac{\partial}{\partial v}\right), Z\right\rangle - \left\langle B\left(\frac{\partial}{\partial v}\right), \nabla_{\frac{\partial}{\partial u}} Z\right\rangle$$

which can be rewritten

$$\frac{\partial}{\partial u}\ b\left(\frac{\partial}{\partial v}, Z\right) = -\left\langle \nabla_{\frac{\partial}{\partial u}} B\left(\frac{\partial}{\partial v}\right), Z\right\rangle + b\left(\frac{\partial}{\partial v}, \nabla_{\frac{\partial}{\partial u}} Z\right)$$

and so

$$\frac{\partial}{\partial u}\ b\left(\frac{\partial}{\partial v}, Z\right) - b\left(\frac{\partial}{\partial v}, \nabla_{\frac{\partial}{\partial u}} Z\right) = -\left\langle \nabla_{\frac{\partial}{\partial u}} B\left(\frac{\partial}{\partial v}\right), Z \right\rangle.$$

There is a similar formula obtained by interchanging u and v, and the result clearly follows. QED

Remark: From here, it is easy to get the classical equations. We have been writing the second fundamental form as $\sum_{ij} b_{ij} X_i Y_j$. Let us adopt the classical language.

Definition 23 *Define*

$$e = b_{11} \qquad f = b_{12} \qquad g = b_{22}$$

Then

$$b(X, X) = e\ X_1^2 + 2f X_1 X_2 + g X_2^2.$$

Theorem 54 *The second and third fundamental surface equations are equivalent to the two equations below, known as the equations of Codazzi-Mainardi.*

$$\frac{\partial e}{\partial v} - \frac{\partial f}{\partial u} = e\Gamma_{12}^1 + f\left(\Gamma_{12}^2 - \Gamma_{11}^1\right) - g\Gamma_{11}^2$$
$$\frac{\partial f}{\partial v} - \frac{\partial g}{\partial u} = e\Gamma_{22}^1 + f\left(\Gamma_{22}^2 - \Gamma_{12}^1\right) - g\Gamma_{12}^2$$

Proof: Let $Z = \frac{\partial}{\partial u}$ in the previous theorem. Then

$$\frac{\partial}{\partial u}\ b\left(\frac{\partial}{\partial v}, Z\right) = \frac{\partial}{\partial u}\ f$$

and

$$\frac{\partial}{\partial v}\ b\left(\frac{\partial}{\partial u}, Z\right) = \frac{\partial}{\partial v}\ e.$$

Also

$$b\left(\frac{\partial}{\partial v}, \nabla_{\frac{\partial}{\partial u}} Z\right) = b\left(\frac{\partial}{\partial v}, \Gamma_{11}^1 \frac{\partial}{\partial u} + \Gamma_{11}^2 \frac{\partial}{\partial v}\right) = f\Gamma_{11}^1 + g\Gamma_{11}^2$$

and

$$b\left(\frac{\partial}{\partial u}, \nabla_{\frac{\partial}{\partial v}} Z\right) = b\left(\frac{\partial}{\partial u}, \Gamma_{12}^1 \frac{\partial}{\partial u} + \Gamma_{12}^2 \frac{\partial}{\partial v}\right) = e\Gamma_{12}^1 + f\Gamma_{12}^2$$

So

$$\frac{\partial}{\partial u}\ b\left(\frac{\partial}{\partial v}, Z\right) - b\left(\frac{\partial}{\partial v}, \nabla_{\frac{\partial}{\partial u}} Z\right) = \frac{\partial}{\partial v}\ b\left(\frac{\partial}{\partial u}, Z\right) - b\left(\frac{\partial}{\partial u}, \nabla_{\frac{\partial}{\partial v}} Z\right)$$

becomes

$$\frac{\partial}{\partial u}\ f - \left(f\Gamma_{11}^1 + g\Gamma_{11}^2\right) = \frac{\partial}{\partial v}\ e - \left(e\Gamma_{12}^1 + f\Gamma_{12}^2\right),$$

which is equivalent to the first equation in the theorem. The second equation can be obtained by setting $Z = \frac{\partial}{\partial v}$. QED.

5.7 Where To Go From Here

There are two possible directions to proceed from the complicated surface equations of this chapter. Mathematicians interested in embedding problems will study the possible choices of b_{ij} for a known geometry g_{ij}. For example, suppose the geometry is Euclidean. We'll later prove that this happens exactly when $\kappa_1 \kappa_2 = 0$. Then we can choose coordinates such that $g_{ij} = \delta_{ij}$ and the interesting question is *how can this geometry be embedded in* R^3 or *what are the surfaces in* R^3 *with Gaussian curvature zero* or *what are the possible* b_{ij} *satisfying the surface equations for* $g_{ij} = \delta_{ij}$*?* By folding a piece of paper without tearing, you'll discover that there are surfaces with Gaussian curvature zero which are not planes, cylinders, or cones.

We will not pursue these questions.

A second direction is to understand the role of Gaussian curvature in the intrinsic geometry of two-dimensional objects. This is the direction Gauss recommended and the direction we will pursue. The surface equations tell us that there is an additional deep geometric invariant, the Gaussian curvature κ, which can be measured by two-dimensional workers. Gauss believed that geometers ignore this invariant at their peril. When the invariant is taken into account, classical theorems going back to Euclid generalize and provide insight into modern developments in mathematics.

Gauss wrote his paper at exactly the moment that Bolyai and Lobachevsky were inventing non-Euclidean geometry. These mathematicians discovered a geometry unlike conventional Euclidean geometry, and yet a geometry which might, for quite plausible reasons, be the correct geometry of the universe. The new geometry had bizarre features:

1. the area of a triangle is completely determined by the three angles of the triangle
2. the area of the entire plane is infinite, but there is a finite bound B such that no triangle has area greater than B
3. the length of a circle is not $2\pi r$
4. the area of a circle is not πr^2
5. perfect rectangles do not exist
6. if two figures are similar, they are congruent.

Much later, Poincare discovered that the Poincare disk is a model for the new geometry. But even before that, Gauss had observed that the features of the new geometry become completely natural once one calculates that its Gaussian curvature is -1. We'll study Gauss's arguments in the remaining chapters of the course.

Chapter 6

The Gauss-Bonnet Theorem

6.1 Introduction

Euclid divided *The Elements* into thirteen books. Each book is a collection of propositions and proofs, with no intermediate explanations. But careful reading reveals a plot.

Book 1 has 48 propositions, divided into five major sections. The first contains preliminary material, including various straightedge and compass constructions. The second, from proposition 8 through proposition 26, is about congruent triangles.

The third section, from proposition 27 through proposition 32, introduces the parallel postulate for the first time, uses this proposition to show that parallel lines cut by a transversal produce equal alternate angles, and culminates in the proof from this result that the sum of the angles of a triangle is 180 degrees.

In the fourth section, Euclid discusses area. He defines figures to be "equal" if they can be decomposed into congruent pieces and uses this notion to prove results which effectively compute the areas of triangles, rectangles, and parallelograms. The section includes propositions from 33 through 43.

The last propositions, from 44 on, lead to the Pythagorian theorem (number 47) and its converse (number 48).

Remark: The intrinsic geometry of surfaces starts with the metric tensor g_{ij}. In some sense, this metric tensor is the Pythagorian theorem in disguise, because we can always choose an orthonormal basis and then the length of $X = X_1 e_1 + X_2 e_2$ is exactly

$$\sqrt{X_1^2 + X_2^2}.$$

Intrinsic geometry provides an answer to the question *if the Pythagorian theorem holds infinitesimally, what are the consequences for geometry?*

Remark: If we read Book 1 of the Elements backward starting with the Pythagorian theorem, we come next to the notion of area. To find the area of a region $\mathcal{R}$ on a surface, we divide the region into parallelograms of size du by dv, find the area of each parallelogram, and add.

Theorem 55 *The area of a region $\mathcal{R}$ on a surface is*

$$\int\int_{\mathcal{R}} \sqrt{g_{11}g_{22} - g_{12}^2}\, dudv$$

Proof: It suffices to show the area of the parallelogram spanned by $X = \frac{\partial}{\partial u}$ and $Y = \frac{\partial}{\partial v}$ is $\sqrt{g_{11}g_{22} - g_{12}^2}$. But $g_{11}g_{22} - g_{12}^2$ equals

$$||X||^2\, ||Y||^2 - \langle X, Y\rangle^2 = ||X||^2\, ||Y||^2\, (1 - \cos^2\theta) = ||X||^2\, ||Y||^2 \sin^2\theta$$

which is the square of the area because

$$\text{area} = \text{base} \times \text{altitude} = ||X|| \times (||Y|| \sin\theta)$$

QED.

More generally, we can integrate an arbitrary function f over $\mathcal{R}$ by dividing the region into parallelograms, multiplying the value of f on each parallelogram by its area, and adding.

Theorem 56 *The integral of $f(u, v)$ over a region $\mathcal{R}$ on a surface is*

$$\int\int_{\mathcal{R}} f(u, v)\sqrt{g_{11}g_{22} - g_{12}^2}\, dudv$$

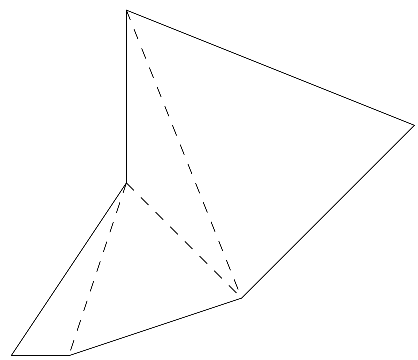
Proof: This follows from the previous theorem and the intuitive definition of the integral.

Remark: The two previous theorems can be taken as definitions if the reader desires.

6.2 Polygons in the Plane

Continuing to read Euclid's book backward, we come next to the section on the sum of the angles of a triangle. Euclid's main result takes a spectacular form in differential geometry, as we'll show.

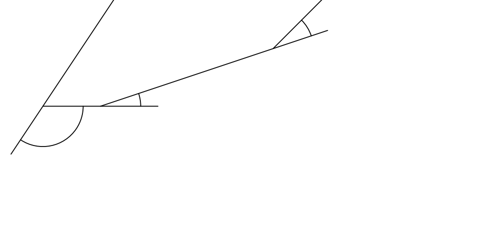
According to Proposition 32 of Euclid, the sum of the angles of a triangle is π. We easily generalize this to arbitrary rectilinear figures. If such a figure has n sides, the sum of the angles is $(n-2)\pi$, as can be seen by dividing the figure into triangles.



We get an equivalent, but easier to remember, theorem by replacing interior angles α_i with exterior angles θ_i, as in the figure below. Then $\sum \theta_i = \sum(\pi - \alpha_i) = n\pi - \sum \alpha_i = n\pi - (n-2)\pi = 2\pi$.

Theorem 57 *The sum of the exterior angles of a rectilinear figure is 2π.*

It is easy to see that this result is reasonable. Travel along the curve counterclockwise. At corners, turn to face the next portion of the curve. You'll have turned completely around by the end of the trip, so the total amount of turning will be 2π.

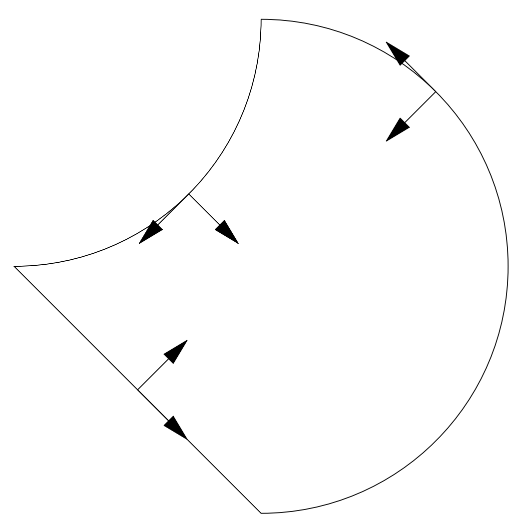


When we measure exterior angles, we must keep track of orientation. In the above figure, notice that most turns are counterclockwise, but one is clockwise. From now on, we assume that the plane has been given the standard right-handed orientation in which angles are measured counterclockwise. Later when we consider surfaces with a predetermined orientation, we insist that local coordinates be chosen so the *uv*-plane is oriented in this standard manner.

Next imagine that our region has curved sides, as in the picture below. At each point on one of these sides, choose a unit tangent vector T and a unit normal vector N. We insist that N be obtained from T by rotating counterclockwise. Each boundary piece has curvature $\kappa(s)$; this curvature can be positive or negative because we have chosen a particular direction for N. We will later prove the following wonderful generalization of Euler's theorem:

Theorem 58 *Suppose $\mathcal{R}$ is a region in the plane whose boundary consists of one counterclockwise curve, possibly with curved sides and corners. Then*

$$\sum_{corners} \theta_i + \sum_{sides} \int_{\gamma_i} \kappa(s) = 2\pi.$$



Example 1: Consider a circle of radius R. This circle has no corners and its curvature is $\frac{1}{R}$. So

$$\int_{\gamma} \kappa(s) = \frac{1}{R} \text{ (length of circle)} = 2\pi$$

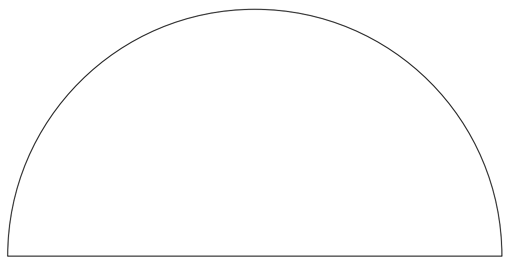
and so

$$(\text{length of circle}) = 2\pi R.$$

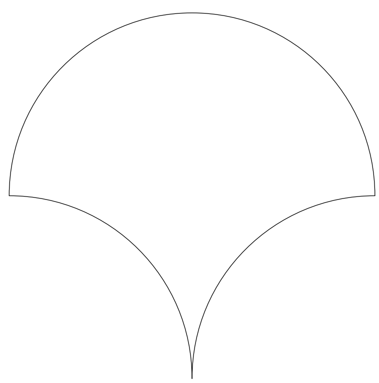
Remark: Consequently, the formula for the length of a circle and the theorem that the sum of the angles of a triangle is π are really special cases of a common result!

Example 2: Consider the semicircle of radius R pictured below. This semicircle has two corners, each of angle $\pi/2$. The previous theorem then yields the following correct result:

$$\sum_{\text{corners}} \theta_i + \sum_{\text{sides}} \int_{\gamma_i} \kappa(s) = \left(\frac{\pi}{2} + \frac{\pi}{2}\right) + \left(\frac{1}{R}(\pi R)\right) = 2\pi.$$



Example 3: Consider the figure below, whose sides are portions of circles of radius R.



The boundary of this figure contains three angles. The exterior angles at the sides are $\pi/2$. The exterior angle at the bottom is π; indeed if the angle at the bottom were slightly less sharp then the exterior angle would clearly be measured counterclockwise and equal slightly less than π.

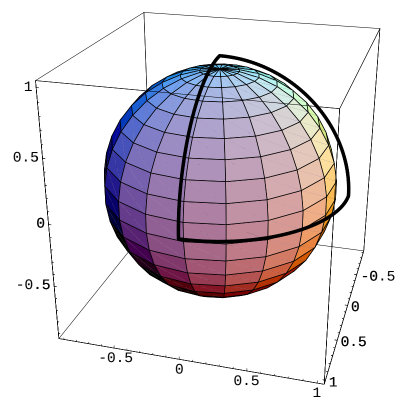
The semicircle at the top has curvature $\frac{1}{R}$, but the two bottom pieces have curvature $-\frac{1}{R}$ because of our orientation convention. So

$$\sum_{\text{corners}} \theta_i + \sum_{\text{sides}} \int_{\gamma_i} \kappa(s) = \left(\frac{\pi}{2} + \frac{\pi}{2} + \pi\right) + \left(\frac{1}{R}(\pi R) - \frac{1}{R}\frac{\pi R}{2} - \frac{1}{R}\frac{\pi R}{2}\right)$$

Again, we get the answer 2π.

6.3 The Gauss-Bonnet Theorem

The previous theorem is certainly false if we replace the plane with a curved surface and replace the curvature of the sides with geodesic curvature. For example, the sides of the region shown below on the sphere are geodesics with geodesic curvature zero, but the sum of the three corner exterior angles is $\frac{3\pi}{2}$ rather than 2π.



Gauss, and later Bonnet, discovered that we can correct the error by adding a single term to the equation — a term containing the Gaussian curvature $\kappa_1\kappa_2$. Gauss proved this theorem when the sides are geodesics, and Bonnet extended it to regions with curved sides. By any measure, the resulting theorem is among the greatest results in mathematics:

Theorem 59 (Gauss-Bonnet) *Let $\mathcal{R}$ be a simply-connected region on a surface, whose boundary consists of a finite number of curves meeting at corners. Then*

$$\sum_{\textit{corners}} \theta_i + \sum_{\textit{sides}} \int_{\gamma_i} \kappa_g(s) + \int\int_{\mathcal{R}} \kappa_1\kappa_2 = 2\pi.$$

Example: Suppose the sphere in the previous example has radius R. The Gaussian curvature is $\frac{1}{R} \cdot \frac{1}{R}$ and the area of the triangle is $\frac{4\pi R^2}{8}$, so

$$\int\int_{\mathcal{R}} \kappa_1 \kappa_2 = \frac{1}{R^2} \frac{4\pi R^2}{8} = \frac{\pi}{2}$$

and the sum of the exterior angles plus this extra term is

$$\left(\frac{\pi}{2} + \frac{\pi}{2} + \frac{\pi}{2}\right) + \frac{\pi}{2} = 2\pi.$$

Remark: The proof of the Gauss-Bonnet theorem is very beautiful. But contrary to our usual custom, we will place it at the end of this chapter after discussing some of the remarkable consequences of the theorem.

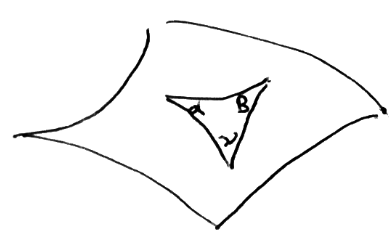
6.4 Initial Applications

We'll discuss some results here whose proofs are immediate and can be left to the reader.

The Gauss-Bonnet theorem explains how two dimensional workers might discover that the Gaussian curvature is nonzero, and compute its value. Namely:

Theorem 60 *Let T be a small triangle on which $\kappa_1\kappa_2$ is essentially constant. Assume the sides of this triangle are geodesics and the angles of the triangle are α, β, and γ. Then*

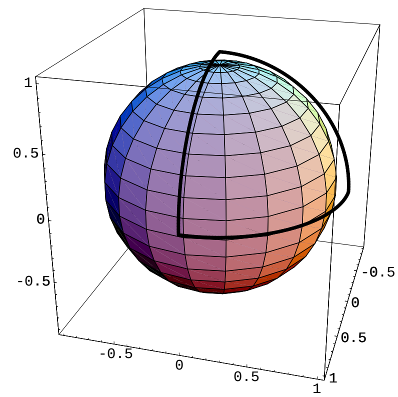
$$\kappa_1 \kappa_2 \sim \frac{(\alpha + \beta + \gamma) - \pi}{(\textit{area of the triangle})}$$



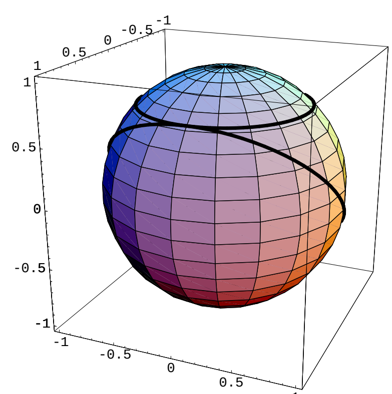
In particular, the sphere of radius one has Gaussian curvature one, so we obtain

Theorem 61 *Let T be a triangle on the sphere of radius one whose sides are great circles. Call the angles of the triangle α, β, and γ. Then*

$$\textit{area of triangle} = \alpha + \beta + \gamma - \pi.$$



Remark: At the end of chapter four, we computed the geodesic curvature of a latitude line on the sphere of radius one. The Gauss-Bonnet theorem gives us an alternate way to do the calculation.



Apply the Gauss-Bonnet theorem to the cap at the top of the sphere. All exterior angles are zero, so

$$\kappa_g \text{ (length of latitude)} + \kappa_1 \kappa_2 \text{ (area of cap)} = 2\pi$$

But the length of the latitude is $2\pi \times \text{radius} = 2\pi \sin\phi$. Also $\kappa_1\kappa_2 = 1$, so we conclude that

$$\kappa_g = \frac{2\pi - (\text{area of cap})}{2\pi \sin\phi}.$$

In section 4.10 we discovered that $g_{11} = \sin^2\phi, g_{12} = 0$, and $g_{22} = 1$. So $\sqrt{g_{11}g_{22} - g_{12}^2}$ equals $\sin\phi$ and the area of the cap is

$$\int_0^{2\pi} \int_0^{\phi} \sin\phi \; d\phi d\theta = 2\pi \; (1 - \cos\phi).$$

Therefore

$$\kappa_g = \frac{2\pi - 2\pi(1 - \cos\phi)}{2\pi \sin\phi} = \frac{\cos\phi}{\sin\phi}.$$

6.5 Non-Euclidean Geometry

In the early part of the nineteenth century, Bolyai and Lobachevsky independently invented non-Euclidean geometry. They replaced Euclid's parallel postulate with a postulate which asserts that more than one line can be drawn through a point parallel to a given line, and investigated the consequences. Bolyai and Lobachevsky worked geometrically; their proofs look very much like Euclidean proofs but their theorems are quite different.

It is unfortunate that the terminology "non Euclidean geometry" were chosen to describe the new geometry, since these words imply that *any* geometry diverging from Euclid is non-Euclidean. Actually, Bolyai and Lobachevsky discovered that there is only one geometric object which satisfies their axiom and the remaining Euclidean axioms.

Gauss seems to have independently studied the geometry, and he rapidly realized that the Gaussian curvature of such a surface would be -1. However, no surface studied in the nineteenth century modeled the complete geometric object. Each candidate surface described only a small piece of the non-Euclidean plane.

The situation changed dramatically when Poincare invented the Poincare disk at the end of the century. This disk is not a surface, but *it is a model* for the complete non-Euclidean plane, and permits us to study non-Euclidean geometry using analytic techniques instead of geometric techniques.

Look back at sections 2.13 and 5.5. These sections sketch proofs of several important non-Euclidean results which we will use here. We summarize them as follows

Theorem 62 *The Poincare disk satisfies:*

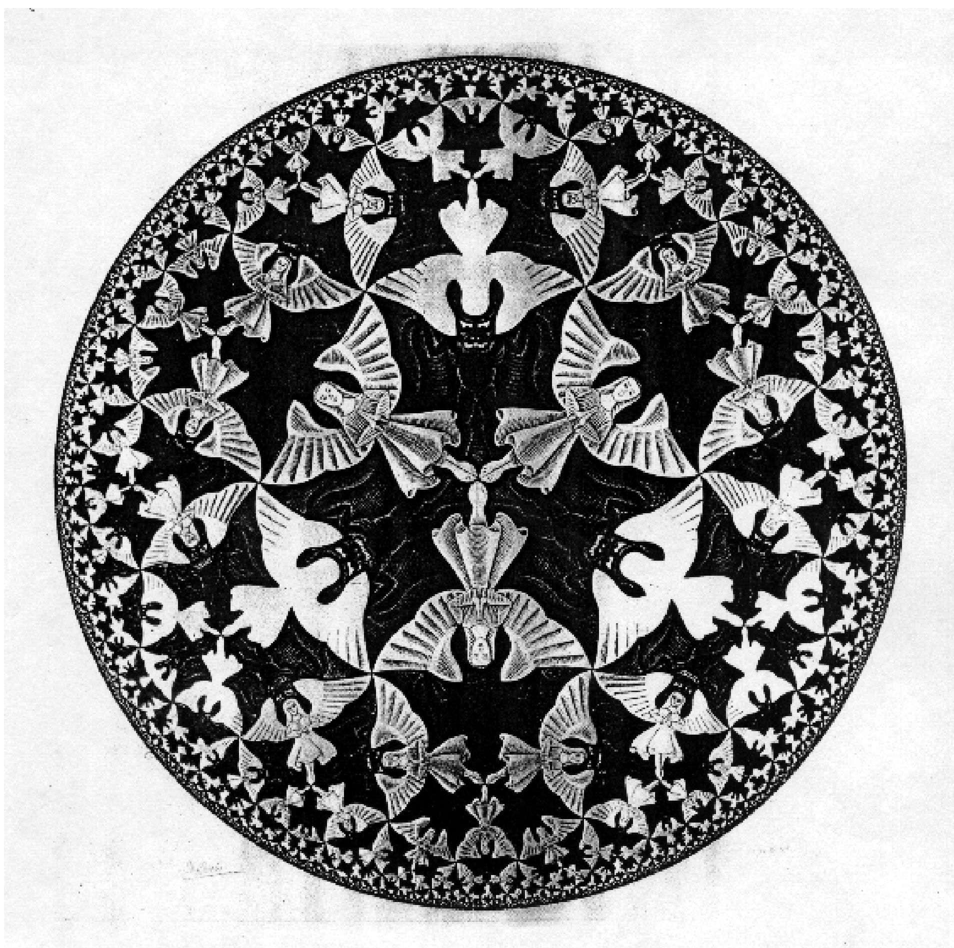
1. *Geodesics in the disk are straight lines through the origin or circles which meet the boundary at ninety degrees.*

2. *The Gaussian curvature is* -1.

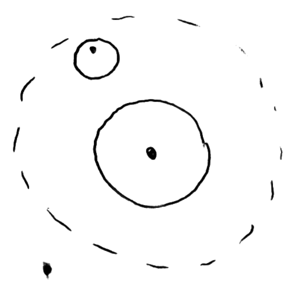
3. *All points of the disk look the same. If* p *and* q *are points, there is a one-to-one, onto map from the disk to itself which preserves all distances, angles, and geodesics, and maps* p *to* q.

4. *All directions on the disk look the same. If* $\{e_1, e_2\}$ *and* $\{f_1, f_2\}$ *are oriented bases of tangent vectors at a point* p, *there is a one-to-one, onto map from the disk to itself which preserves all distances, angles, and geodesics, and maps* e_1 *to* f_1 *and* e_2 *to* f_2.

5. *A non-Euclidean circle looks like a Euclidean circle in the model, except that its center is at an unexpected place.*



Proof: The proofs for most of these results were sketched earlier. The last result can be proved in the following way. Non-Euclidean circles centered at the origin are clearly Euclidean circles (although the non-Euclidean radius is not the same as the Euclidean radius). Suppose p is not the origin and we wish to study circles centered at p. There is an isometry mapping the origin to p. This map preserves distances, and so maps non-Euclidean circles to other non-Euclidean circles. Formulas for these isometries are given in section 2.13. By Lemma 3 in that section, these isometries map Euclidean lines and circles to Euclidean lines and circles. A non-Euclidean circle at the origin is a Euclidean circle, so its image under the isometry, which is known to be a non-Euclidean circle, is also a Euclidean circle. The result follows. QED.



Theorem 63 *Suppose a worker stands at a point p in the non-Euclidean plane. Near p the geometry is almost Euclidean; discrepencies arise only when the worker moves away from p. The geodesic starting at p in the direction X continues forever and reaches distances arbitrarily far away. Every point of the plane can be seen by looking out in exactly one direction. The topology of the non-Euclidean plane is exactly the same as the topology of the Euclidean plane. In all of these respects, non-Euclidean geometry works exactly like Euclidean geometry, with no surprises.*

Proof: The theorem follows from the philosophy of differential geometry, from the obvious homeomorphism carrying the open unit disk to the plane, and from the fact proved earlier that distances to the boundary along straight lines from the origin are infinite. QED.

We come next to the surprising results of the subject. Below are analytic proofs of some of these amazing results, proved synthetically by Bolyai and Lobachevsky:

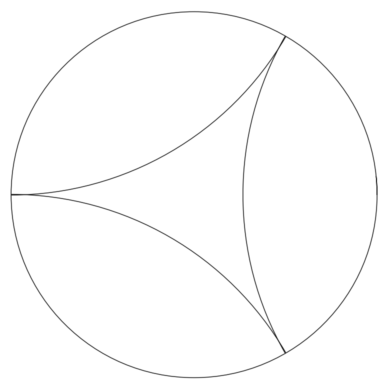
Theorem 64 *In non-Euclidean geometry*

1. *The sum of the angles of a triangle is less than π.*
2. *The area of a triangle with angles α, β, γ is*

$$\pi - (\alpha + \beta + \gamma)$$

3. *The area of a triangle can never be larger than π.*
4. *Triangles exist which have area arbitrarily close to pi.*

Proof: This is an immediate consequence of Gauss-Bonnet together with $\kappa_1\kappa_2 = -1$. The last result is true because the following "triangle" has all three angles equal to zero. In fact this triangle has sides of infinite length and so is illegal, but clearly we can shrink it a little and get an honest triangle with $\alpha + \beta + \gamma$ as small as we like. QED.



Theorem 65 *In non-Euclidean geometry*

1. *Squares with four right angles do not exist.*
2. *Rectangles with four right angles do not exist.*
3. *If a quadrilaterial has angles $\alpha, \beta, \gamma, \delta$, then the area of the quadrilaterial is*

$$2\pi - (\alpha + \beta + \gamma + \delta)$$

4. *The area of a quadrilateral can never be larger than 2π.*
5. *There are quadrilaterials which have area arbitrarily close to 2π.*

Proof: Exactly as before. QED.

Theorem 66 *In non-Euclidean geometry*

1. *The area of a figure with n sides and angles $\alpha_1, \ldots, \alpha_n$ is*

$$(n-2)\pi - (\alpha_1 + \ldots + \alpha_n)$$

2. *The area of a figure with n sides can never be larger than $(n-2)\pi$.*
3. *There are figures with n sides which have area arbitrarily close to $(n-2)\pi$.*

Proof: As before. QED.

Theorem 67 *In non-Euclidean geometry*

1. *The area of the plane is infinite.*
2. *The area of a circle can be arbitrarily large.*
3. *Farmers in non-Euclidean geometry insisting on rectangular fields are restricted to fields with area below a fixed bound. But farmers willing to use circular fields can form fields of arbitrary area.*
4. *The circumference of a circle of radius R is*

$$C = 2\pi \sinh R = 2\pi \left(R + \frac{R^3}{3!} + \frac{R^5}{5!} + \ldots \right)$$

This number is approximately $2\pi R$ for small R.

5. *The area of a circle of radius R is*

$$A = 4\pi \ \sinh^2(R/2) = 2\pi\,(\cosh R - 1) = 2\pi \left(\frac{R^2}{2!} + \frac{R^4}{4!} + \ldots \right)$$

This number is approximately πR^2 for small R.

6. *The geodesic curvature of a circle of radius R is*

$$\kappa_g = \frac{\cosh R}{\sinh R} = \frac{1}{R}\left(1 + \frac{R^2}{6} + \ldots\right)$$

This number is approximately $\frac{1}{R}$ for small R.

Proof: Recall that

$$ds^2 = \frac{4(dx^2 + dy^2)}{((1 - (x^2 + y^2))^2}.$$

It suffices to prove the last three results. Since any circle can be mapped to any other by an isometry, it suffices to study circles about the origin. Such a circle looks like a Euclidean circle of Euclidian radius r. We shall compute its non-Euclidean radius R. The curve $\gamma(t) = (t, 0)$ for $0 \le t \le r$ is a radial curve and its non-Euclidean length is

$$R = \int_0^r \frac{2\ dt}{1 - t^2} = \int_0^r \left(\frac{1}{1-t} + \frac{1}{1+t} \right)\ dt = \ln \left(\frac{1+r}{1-r} \right)$$

Solving this equation for r gives

$$r = \frac{e^R - 1}{e^R + 1}.$$

Next we find the non-Euclidean circumference of this circle. The circle can be parameterized as $(r \cos t, r \sin t)$, so its length is

$$\int_0^{2\pi} \frac{2\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\ dt}{(1 - (x^2 + y^2))} = \int_0^{2\pi} \frac{2r}{1 - r^2}\ dt = \frac{2r}{1 - r^2}\ 2\pi.$$

Substitution of our previous formula for r in terms of R gives

$$\text{length} = 2\pi \cdot 2 \left(\frac{e^R - 1}{e^R + 1} \right) \Big/ \left(1 - \left(\frac{e^R - 1}{e^R + 1} \right)^2 \right) = 2\pi \left(\frac{e^R - e^{-R}}{2} \right) = 2\pi \sinh R.$$

The non-Euclidean area of the disk equals:

$$\int\int \frac{2dx}{1 - (x^2 + y^2)} \times \frac{2dy}{1 - (x^2 + y^2)}.$$

It is convenient to integrate this using polar coordinates. (Polar coordinates introduce a momentary clash between the meaning of r as a polar coordinate and the meaning of r as a limit of integration, but we survive!)

In polar coordinates this integral becomes

$$\int_0^{2\pi} \int_0^r \frac{4\ r\ dr d\theta}{(1 - r^2)^2} = 2\pi \int_0^r \left(\frac{1}{(r-1)^2} - \frac{1}{(r+1)^2} \right)\ dr = 2\pi \left[\left(\frac{1}{r+1} - \frac{1}{r-1} \right) - 2 \right]$$

This simplifies to

$$2\pi \frac{2r^2}{1 - r^2}.$$

Substitution of our formula for r in terms of R yields, after a brief simplification,

$$\text{area} = 4\pi \left(\frac{e^{R/2} - e^{-R/2}}{2} \right)^2 = 4\pi \sinh^2 R/2.$$

Recall that $\cosh R = \cosh(R/2+R/2) = \cosh^2(R/2)+\sinh^2(R/2)$. Since $\cosh^2 x - \sinh^2 x = 1$, we obtain $\cosh R = 1 + 2\sinh^2(R/2)$ and the formula for area follows.

To obtain the geodesic curvature of a circle of radius R, we apply the Gauss-Bonnet theorem to this circle. The circle has no exterior angles, so the theorem states that

$$\kappa_g(\text{length of circle}) + (-1)(\text{area of circle}) = 2\pi$$

or

$$\kappa_g(2\pi \sinh R) - 2\pi(\cosh R - 1) = 2\pi.$$

So

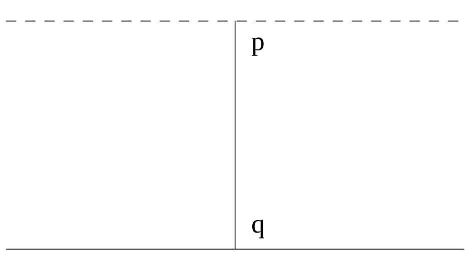
$$\kappa_g = \frac{\cosh R}{\sinh R}.$$

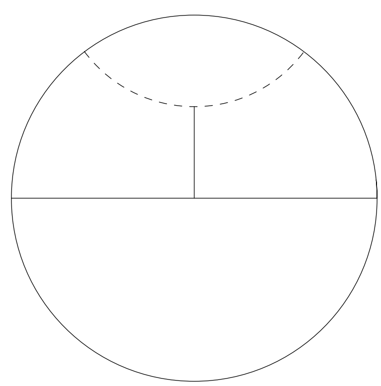
Compare this result with the analogous result on the sphere. QED.

Remark: Finally let us study parallel lines.

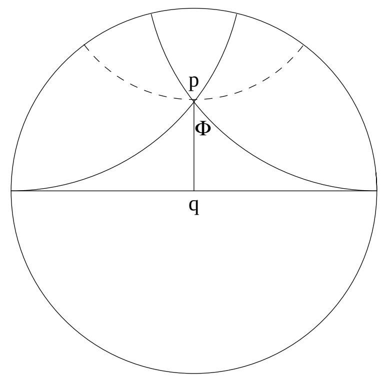
In Euclidean geometry, we often apply the following construction: Given a line l and a point $q \in l$, draw a perpendicular to l at q and extend it to a point p. Draw a perpendicular to this line segment, obtaining a new line parallel to l. See the left picture below.

This construction still works in non-Euclidean geometry. Since the Poincare disk has a large number of isometries, it is enough to study the situation when q is the origin and l is the x-axis. The right picture below shows the non-Euclidean version of the construction.





But in non-Euclidean geometry there are infinitely many other lines through p parallel to l. See the picture below, and notice that there are two particular lines that are just barely parallel. Lobachevsky called these lines the *limiting parallels*.



At p, there is an angle Φ such that all lines draw with angle $0 \leq \theta < \Phi$ meet l, and all lines drawn with angle $\Phi \leq \theta \leq \frac{\pi}{2}$ are parallel to l. The angle Φ depends on the distance d from p to q. If this distance is small, the angle Φ is almost $\frac{\pi}{2}$ and two-dimensional workers may not even notice that more than one parallel can be drawn. But when the distance is larger, the angle Φ approaches zero and it becomes evident that there are many possible parallel lines.

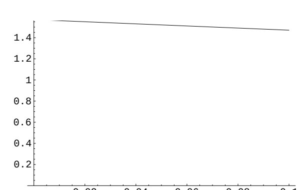
Theorem 68 *The angle Φ is given in terms of the non-Euclidean distance d from q to p by*

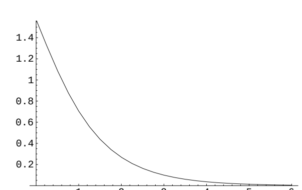
$$\tan\left(\frac{\Phi}{2}\right) = e^{-d}.$$

For small d we have

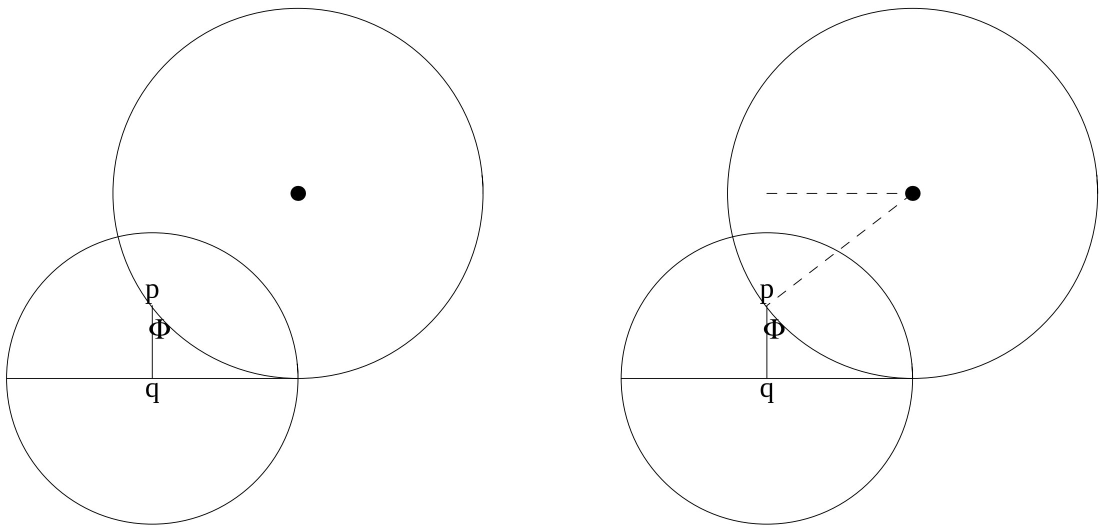
$$\Phi = \frac{\pi}{2} - d + \frac{1}{6}\, d^3 - \frac{1}{24}\, d^5 + \frac{61}{5040}\, d^7 + \ldots$$

For distances smaller than $1/10$, the divergence of Φ from $\frac{\pi}{2}$ is negligible, as shown by the plot below. The divergence becomes significant for d close to 1, and the angle Φ is so small for $6 \leq d \leq \infty$ that non-parallel lines are rare.





Proof: The picture on the left below shows that the center of the circle which gives the limiting parallel in the Poincare model has x-coordinate one (since it meets the boundary of the model perpendicularly at $(1, 0)$). Call the radius of this circle a.



Draw the two additional dotted lines indicated on the right, and notice that these lines meet at an angle Φ because each dotted line is perpendicular to a side of the original angle Φ. The side opposite this new angle is the radius of the large circle minus the Euclidean distance from q to p. If this Euclidean distance is called r, then

$$\sin\Phi = \frac{a-r}{a}.$$

By the Pythagorian theorem applied to the triangle with dotted sides, we have $(a-r)^2+1^2 = a^2$ and so $a = \frac{r^2+1}{2r}$. Consequently

$$\sin\Phi = \frac{1-r^2}{1+r^2}.$$

Then

$$\cos^2\Phi = 1-\sin^2\Phi = \frac{(1+r^2)^2-(1-r^2)^2}{(1+r^2)^2} = \frac{4r^2}{(1+r^2)^2} = \left(\frac{2r}{1+r^2}\right)^2$$

and so

$$\tan\frac{\Phi}{2} = \frac{\sin\Phi}{1+\cos\Phi} = \frac{(1-r^2)/(1+r^2)}{(1+2r/(1+r^2))} = \frac{(1-r)\ (1+r)}{(1+r)\ (1+r)} = \frac{1-r}{1+r}$$

On page 151 we discovered that $d = \ln\left(\frac{1+r}{1-r}\right)$. Therefore

$$\tan\frac{\Phi}{2} = \frac{1-r}{1+r} = e^{-d}.$$

QED.

6.6 Book I of Euclid

In the first section of this chapter, we compared differential geometry to Book I of Euclid. There is one portion of Euclid which we did not discuss. That is the section on congruent triangles, starting with proposition 8.

The first proposition in this section is the familiar side-angle-side congruence theorem. Euclid proves this proposition by *superposition.* He tells us to move the first triangle until the two sides and included angle of the moved triangle lie on top of the corresponding sides and angle of the second triangle. Then, he says, the remaining side and angles clearly also match.

It is easy to see that this proof fails for surfaces. For instance, draw a right angle on the rim of a doughnut where the Gaussian curvature is positive. Extend each side of the angle by a small distance d, forming a small triangle. Since the curvature is positive, the sum of the remaining angles will be greater than $\frac{\pi}{2}$. Draw the same right angle and sides on the inside rim of the doughnut where the Gaussian curvature is negative. Then the sum of the remaining angles will be smaller than $\frac{\pi}{2}$. So side-angle-side fails on a doughnut.

The problem is that there is no isometry from a doughnut to itself carrying the outside triangle to the inside one. So Euclid's superposition proof does not work.

Curiously, Euclid doesn't use superposition again, although many other congruence theorems could be proved that way. Euclid probably was unhappy with his proof of Proposition 8, and later authors took Proposition 8 on side-angle-side to be another axiom. The modern approach is to replace this axiom with the principle of superposition itself. From this point of view, propositions 8 through 26 of Euclid follow from requiring that the surface has many isometries, where

Definition 24 *We say a surface has* sufficiently many isometries *if*

1. *whenever p and q are points of the surface, there is an isometry of the surface which carries p to q*
2. *whenever $\{e_1, e_2\}$ and $\{f_1, f_2\}$ are orthonormal bases of tangent vectors are p, there is an isometry carrying e_1 to f_1 and e_2 to f_2.*

Remark: It is possible to prove that there are only four surfaces with sufficiently many isometries, up to magnification:

1. the Euclidean plane
2. the sphere
3. the sphere with opposite points identified, usually called projective space
4. the non-Euclidean plane

Thus differential geometry is a completely natural complement to Euclid, and the results we have proved recently are natural theorems in a completion of Euclid's work.

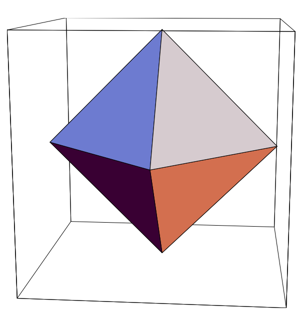
6.7 Compact Surfaces

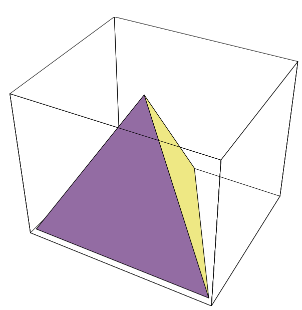
The topologists have proved a glorious theorem about the surfaces of doughnuts with g holes. The theorem was first proved by Euler for surfaces which can be deformed to spheres, and later extended by Poincare to the general case.

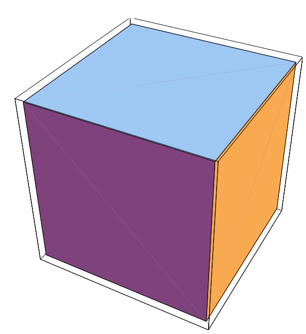
Theorem 69 (Euler-Poincare) *Suppose $\mathcal{S}$ is a surface which can be deformed to a doughnut with g holes. Cut this surface into faces. The sides of these faces are allowed to be curved should be a mixture of triangles, quadrilaterals, 5-sided polygons, etc. Then*

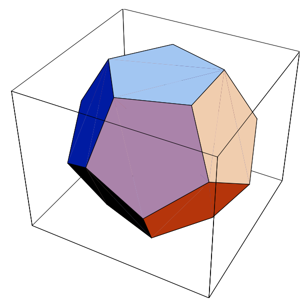
$$(\textit{number of vertices}) - (\textit{number of edges}) + (\textit{number of faces}) = 2 - 2g.$$

Examples: Consider the objects below:



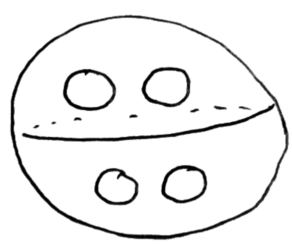




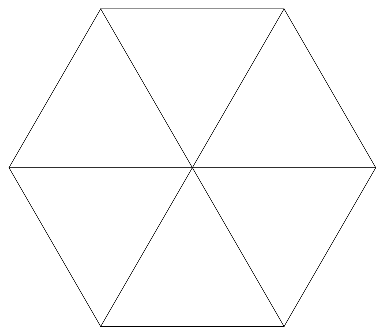


Each of these objects looks like a sphere, so the Euler-Poincare number should equal two. On a cube, there are 8 vertices, 12 edges, and 6 faces, and $8-12+6=2$. On a tetrahedron, there are 4 vertices, 6 edges, and 4 faces, and $4-6+4=2$. Etc.

Proof: Cut the doughnut across each hole, as illustrated on the next page. The cuts should follow edges of the dissection, which is not shown in the pictures. Notice that around the hole there are n vertices and n edges, and after the cut the number of such vertices and such edges doubles. But these terms cancel and the Euler-Poincare number does not change. When we are done, we have an object which looks like a sphere with $2g$ disks removed.

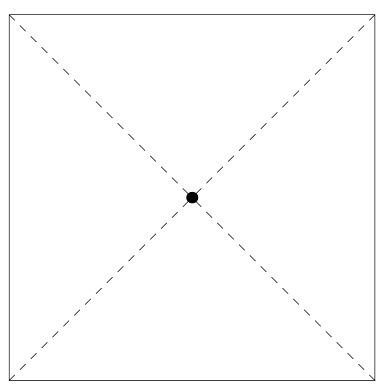


Each hole in the sphere is surrounded by n edges. Fill in each of these holes with a disk cut into n triangles as illustrated below. Notice that each such disk adds 1 vertex, n edges, and n faces to the dissection of the surface, and thus increases the Euler-Poincare number by one. Since there are 2g holes to be filled in, the Euler-Poincare number will increase by $2g$. Thus it was originally $2 - 2g$ if and only if it is 2 after filling in the holes.

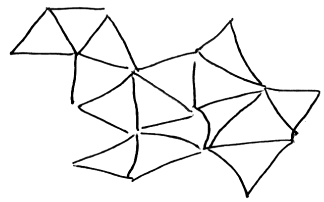


The object now looks like a sphere. So it suffices to prove the theorem in the special case when we have a sphere.

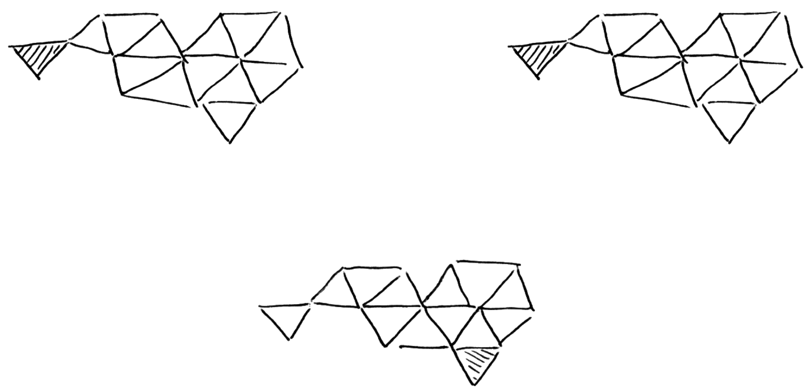
Without loss of generality, we can suppose that each of the faces is a triangle. For if a face has n edges, subdivide it as illustrated below, and notice that the subdivision added n edges, $n - 1$ new faces, and 1 new vertex, so the Euler-Poincare number did not change.



Remove one face and notice that the resulting object can be flattened down into the plane, possibly by greatly distorting the shapes of the faces and edges. The original Euler-Poincare number will have been 2 exactly if the new number is 1, so we want to prove that the Euler-Poincare number of the resulting mass of triangles is one.



Begin removing triangles one by one from the new object. There are three ways to do this, as illustrated on the next page. Note that none of these methods changes the Euler-Poincare number. In the end, we will have one triangle left, and the Euler-Poincare number will be $3 - 3 + 1 = 1$, as desired. QED.



Remark: There is a variant of the Gauss-Bonnet theorem for surfaces. The variant is a consequence of the original Gauss-Bonnet theorem and the above result of Euler-Poincare, as we shall see.

Theorem 70 (Gauss-Bonnet) *Let $\mathcal{S}$ be a surface in the shape of a doughnut with g holes. Then*

$$\frac{1}{2\pi} \int \int_{\mathcal{S}} \kappa_1 \kappa_2 = 2 - 2g.$$

Proof: Cut the surface into triangles using edges which are C^∞ curves, not necessarily geodesics. Compute $\int \int_{\mathcal{S}} \kappa_1 \kappa_2$ by summing this integral over triangles.

$$\frac{1}{2\pi} \int \int_{\mathcal{S}} \kappa_1 \kappa_2 = \frac{1}{2\pi} \sum_{\text{triangles}} \int \int \kappa_1 \kappa_2$$

The Gauss-Bonnet theorem for triangles states that

$$(\pi-\alpha)+(\pi-\beta)+(\pi-\gamma)+\sum_{\text{sides}}\int\kappa_g+\int\int_{\text{triangle}}\kappa_1\kappa_2=2\pi$$

or equivalently

$$\int\int_{\text{triangle}}\kappa_1\kappa_2=(\alpha+\beta+\gamma-\pi)-\sum_{\text{sides}}\int\kappa_g$$

Apply this theorem to obtain

$$\frac{1}{2\pi}\int\int_{\mathcal{S}}\kappa_1\kappa_2=\frac{1}{2\pi}\sum_{\text{triangles}}\left(\alpha+\beta+\gamma-\pi-\sum_{\text{sides}}\int\kappa_g\right)$$

However, the geodesic curvature must be computed using normals which point into the triangle. Each edge of the above decomposition appears twice, bounding two triangles. In one appearance the normal points one way and in the other it points the other way, so the integrals involving κ_g cancel.

We are left with

$$\frac{1}{2\pi}\int\int_{\mathcal{S}}\kappa_1\kappa_2=\frac{1}{2\pi}\sum_{\text{triangles}}(\alpha+\beta+\gamma-\pi)$$

The sum over triangle angles is supposed to be computed by summing the angles of each triangle, and then summing these numbers over triangles. But it could also be computed by summing the angles which meet in a particular vertex, and then summing over vertices. When computed this way, the sum is 2π(number of vertices). The sum over triangles of π gives π(number of triangles). So the above number equals

$$(\text{number of vertices})-\frac{1}{2}(\text{number of triangles})$$

Each triangle has three edges, so the total number of edges is

$$3\ (\text{number of triangles}).$$

But this counts each edge twice because an edge bounds two triangles. So

$$\frac{3}{2}\ (\text{number of triangles})=(\text{number of edges}).$$

Hence

$$\frac{1}{2}\ (\text{number of triangles})=(\text{number of edges})-(\text{number of triangles}).$$

So $\frac{1}{2\pi}\int\int_{\mathcal{S}} \kappa_1\kappa_2$, which was earlier proved to equal

$$\text{(number of vertices)} - \frac{1}{2}\text{(number of triangles)},$$

also equals

$$\text{(number of vertices)} - \text{(number of edges)} + \text{(number of triangles)} = 2 - 2g.$$

QED.

Theorem 71 *Every compact surface in R^3 contains points where the Gaussian curvature is positive. Every compact surface in R^3 except the sphere contains points where the Gaussian curvature is negative.*

Proof: If a surface is compact, there is a point p on the surface whose distance from the origin is a maximum. Consequently the entire surface is inside the sphere of radius $R = ||p||$, and this sphere touches the surface at p. It follows that the principal curvatures κ_1 and κ_2 at p must be at least $\frac{1}{R}$ and have the same sign. So the Gaussian curvature is positive at p.

By the Gauss-Bonnet theorem, $\int\int_{\mathcal{S}} \kappa_1\kappa_2 = 2 - 2g$; this number is zero or negative unless $g = 0$ and the surface is topologically a sphere. Since $\kappa_1\kappa_2$ is continuous and sometimes positive, it must also be sometimes negative. QED.

6.8 Constant Curvature

We return to Euclid for a final time. Euclid built geometry on a small number of axioms. Once we have the machinery of differential geometry, we can replace his axioms by a series of equivalent but more precise axioms. For instance, Euclid assumes that we can draw straight lines and can determine whether line segments are congruent and whether angles are congruent. The analogous modern assumption is that a geometry is a two-dimensional surface (not necessarily in R^3) and a metric tensor g_{ij} on this surface. Given this information, we can measure lengths of curves and angles, and thus determine congruence for such objects. We can draw geodesics, and thus speak of straight lines in Euclid's sense.

A long section of Book 1 of Euclid is about congruent triangles. In section 6.6, we argued that Euclid's proof of the congruence theorems ultimately relies on a superposition argument whose modern equivalent is the existence of a large number of surface isometries. We want to expand on that idea.

In some sense, the congruence theorems (and analogous axioms requiring a large number of isometries) are at heart assertions that space is homogeneous — that geometry near one point p is the same as geometry near another point q. This axiom is the opposite of the ancient belief that the earth is the center of the universe, or that a sacred city (Rome, Jerusalem, or Mecca) is the center of the world. According to the congruence-isometry axiom, all points are on an equal footing.

We'd like to convert this philosophy into mathematics. We are going to ignore the unfortunate discovery that space is curved due to gravity by different amounts at different locations and thus is not homogeneous.

According to the isometry axiom, whenever p and q are points, there is a one-to-one and onto map from the surface to itself preserving all distances, angles, and geodesics, and carrying p to q. This requirement is very restrictive and somewhat implausible as an axiom. For instance, if we believe that geometry is the same near the earth and near the star Alpha Centura, the axiom requires us to move the earth to Alpha Centura *and simultaneously move all of the stars in the universe to new positions!*.

A better axiom would require local motion only, along the lines of the following definition:

Definition 25 *We say that a surface is* locally homogeneous *if whenever p and q are points, there are open neighborhoods $\mathcal{U}$ and $\mathcal{V}$ of p and q and a one-to-one, onto C^∞ map from $\mathcal{U}$ to $\mathcal{V}$, with C^∞ inverse, preserving the metric g_{ij}.*

Remark: In particular, all geometric notions are preserved by the local map, so the Gaussian curvature will be the same at p and q. Since p and q are arbitrary, the Gaussian curvature will be constant on a locally homogeneous surface.

In the final chapter of these notes, we will prove that conversely if the Gaussian curvature is constant, then any pair of points p and q have neighborhoods which are isometric. So a useful replacement for Euclid's congruence axiom is the requirement that the surface have constant curvature.

So much for "philosophy." In the next section, we convert this philosophy into interesting mathematics.

6.9 Surfaces of Constant Curvature

Suppose $\mathcal{S}$ is a surface with constant curvature and metric tensor g_{ij}. Let us *magnify* this surface by a factor $m > 0$. To do so, multiply all lengths by m and leave all angles unchanged. It is easy to see that this can get done by multiplying g_{ij} by m^2.

Theorem 72 *If all distances on a surface are multiplied by m, then the Gaussian curvature is multiplied by $\frac{1}{m^2}$.*

Proof: Turn back to the formula for Γ_{ij}^k on page 52. Notice that the terms of g^{-1} are multiplied by $\frac{1}{m^2}$ and thus Γ_{ij}^k is unchanged.

Turn to the formula for the curvature tensor on page 126. Notice that this tensor is multiplied by m^2. Turn finally to the formula for Gaussian curvature in terms of the curvature tensor on page 127. Notice that R_{1212} is multiplied by m^2 and $\det(g_{ij})$ is multiplied by m^4 and consequently $\kappa_1\kappa_2$ is multiplied by $\frac{1}{m^2}$. QED.

Remark: It follows that if $\mathcal{S}$ is a surface of constant curvature, we can magnify the surface appropriately so its curvature is -1, 0, or 1. In the future, we always assume that this has been done.

Theorem 73 *Suppose $\mathcal{S}$ is a compact oriented surface, and thus a doughnut with g holes. We do not assume that this surface is embedded in R^3.*

1. *If the surface has a metric of positive constant curvature, then $g = 0$ and the surface is a sphere. If the surface is magnified to make its curvature 1, then its area is 4π.*
2. *If the surface has a metric of zero constant curvature, then $g = 1$ and the surface is a doughnut.*
3. *If the surface has a metric of constant negative curvature, then $g \geq 2$ and the surface is a doughnut with at least two holes. If the surface is magnified to make its curvature -1, then its area is $4\pi(g-1)$.*

Proof: This follows immediately from the Gauss-Bonnet theorem. For instance, if the curvature is 1, then

$$\frac{1}{2\pi}\int\int \kappa_1\kappa_2 = \frac{1}{2\pi}\ (\text{area of surface}) = 2-2g,$$

so $2-2g>0$ and thus $g=0$. Then

$$(\text{area of surface}) = (2\pi)\ (2-2g) = 4\pi.$$

The remaining cases are proved in a similar fashion. QED.

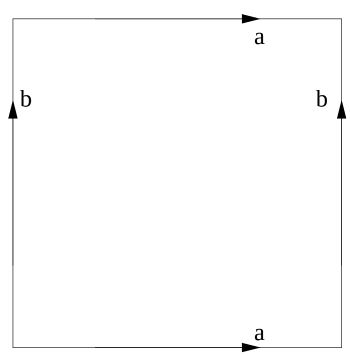
Remark: The amazing fact is that the converse is true!

Theorem 74 *Let $\mathcal{S}$ be a compact oriented surface with g holes.*

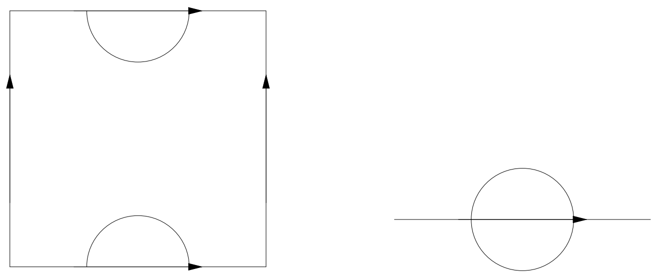
1. *If $g=0$, then $\mathcal{S}$ has a metric with constant curvature 1.*
2. *If $g=1$, then $\mathcal{S}$ has a metric with constant curvature 0.*
3. *If $g\geq 2$, then $\mathcal{S}$ has a metric with constant curvature -1.*

Proof: The first result is clear since the unit sphere has curvature 1.

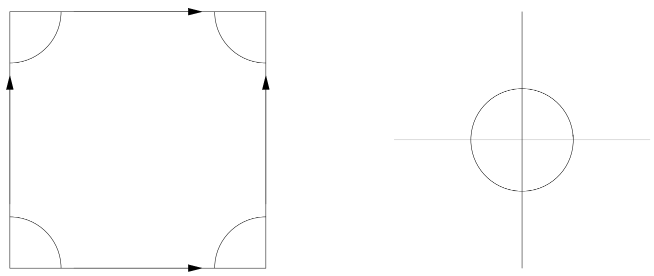
We give two arguments for the torus. Here is the first. Form a torus from a unit square by glueing the top and bottom together to form a cylinder, and then glueing the left and right together to form a doughnut. Give the square the standard flat Euclidean metric.



We must make sure that each point has a Euclidean neighborhood after the glueing is complete. This is clear for interior points. It is clear for boundary points from the following picture.



Finally, all four corners of the square glue to become a single point, and this point has a Euclidean neighborhood because the four corner angles, each of size $\frac{\pi}{2}$, glue together to form a neighborhood with total angle 2π as illustrated below.



Here is a second argument for the torus. The torus can be embedded in R^3 by the map illustrated on page 30:

$$s(\theta, \varphi) = ((R + r\cos\varphi)\cos\theta, (R + r\cos\varphi)\sin\theta, r\sin\varphi)$$

Here $0 \le \varphi \le 2\pi$ and $0 \le \theta \le 2\pi$. However, we can also embedded the torus in R^4, and in a considerably easier manner:

$$s(\theta, \varphi) = (\cos\theta, \sin\theta, \cos\varphi, \sin\varphi) .$$

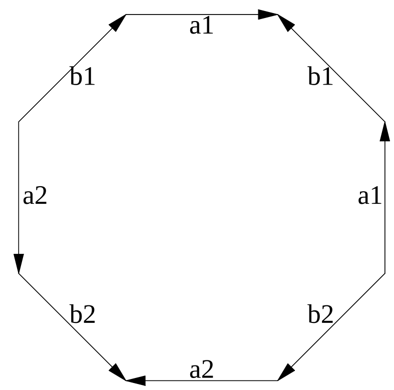
Let g_{ij} be the induced metric. Notice that

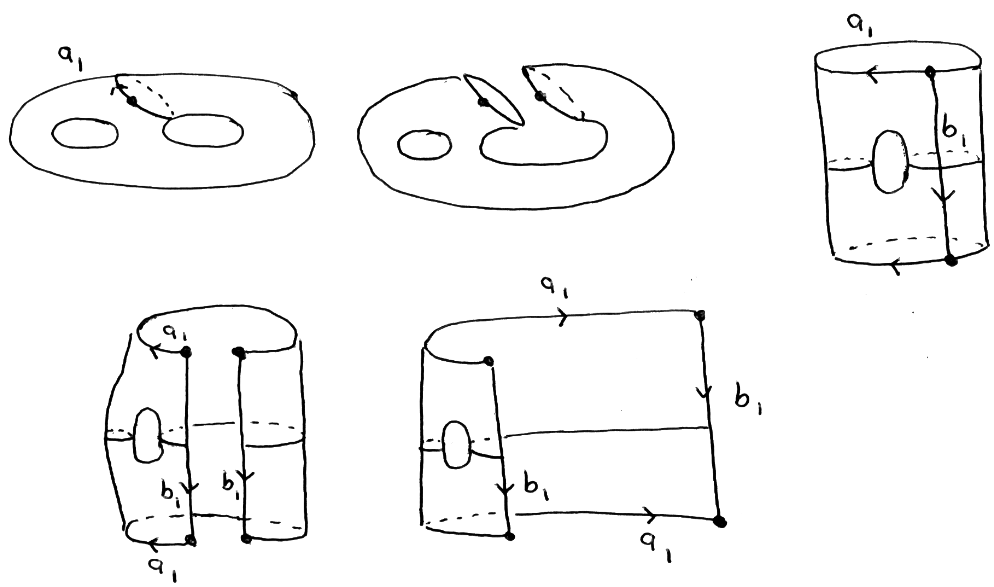
$$\frac{\partial s}{\partial \theta} = (-\sin\theta, \cos\theta, 0, 0) \qquad \frac{\partial s}{\partial \varphi} = (0, 0, -\sin\varphi, \cos\varphi)$$

and so

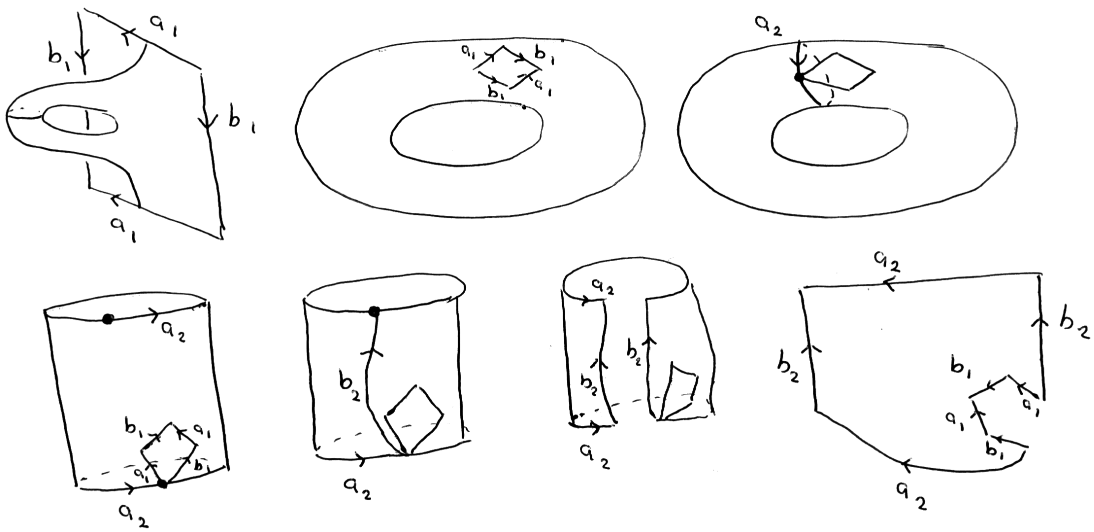
$$g_{11} = \frac{\partial s}{\partial \theta} \cdot \frac{\partial s}{\partial \theta} = 1 \qquad g_{12} = \frac{\partial s}{\partial \theta} \cdot \frac{\partial s}{\partial \varphi} = 0 \qquad g_{22} = \frac{\partial s}{\partial \varphi} \cdot \frac{\partial s}{\partial \varphi} = 1.$$

Finally, here is an argument for surfaces with $g \ge 2$ holes. The topologists have proved that every such surface can be constructed from a regular polygon with 4g sides by identifying corresponding sides as illustrates below. The pictures on the next page show this process for a doughnut with $g = 2$. In particular, all vertices of the polygon glue to the same point in the surface.

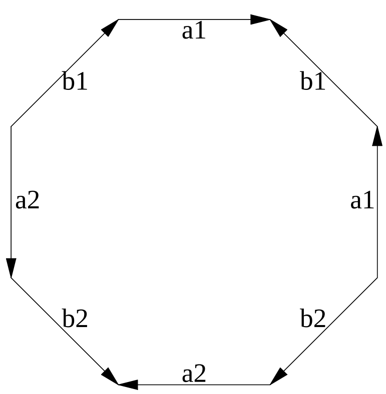




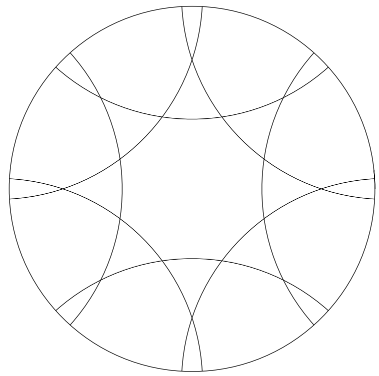
Flatten this object so the remaining doughnut tube sticks out toward us:



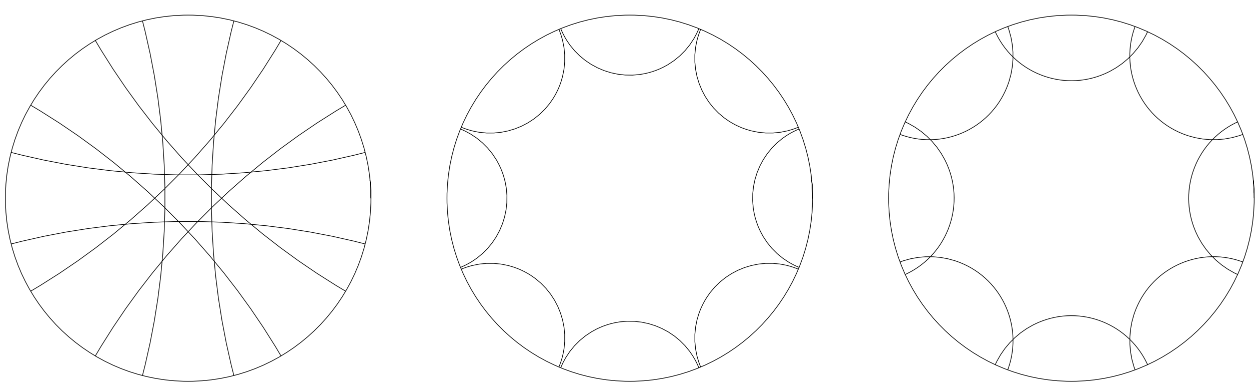
We will attempt to glue the sides of our polygon together so the resulting surface inherits a geometry with constant curvature. Let's first try to do that using flat Euclidean geometry.



After glueing, every point inside the polygon or on the sides has a Euclidean neighborhood as in the earlier torus example. But there is a problem at the point obtained by glueing all vertices of the polygon together because the sum of the interior angles of the polygon at the vertices adds up to more than 2π. To fix this problem, we will construct the polygon in non-Euclidean geometry rather than in the plane. Consider the polygon at the center of the picture on the next page. Each side of this polygon is a geodesic. We construct the polygon so these sides all have the same length. Then after gluing each point which comes from the inside of the polygon and each point which comes from a boundary line has a non-Euclidean neighborhood.



We must check that the vertex angles sum to 2π. We do this by adjusting the size of the polygon. If the polygon is very small as on the left below, the sum of the angles of the polygon will be close to the corresponding sum for Euclidean polygons, which is larger than 2π. If the polygon touches the boundary as in the middle below, then all angles will be zero and the angle sum is zero. Somewhere between these extremes, there is a non-Euclidean polygon whose vertex angle sum is exactly 2π. QED.

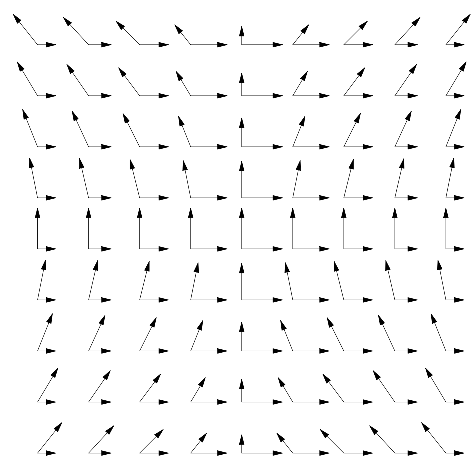


6.10 Frames

This section and the remaining sections of the chapter yield a proof of the Gauss-Bonnet theorem. Several interesting ideas arise along the way.

Definition 26 *A* framing *of a coordinate system* $\mathcal{U}$ *is an assignment to each point* $p \in \mathcal{U}$ *of an orthonormal basis* $\{e_1(p), e_2(p)\}$ *of tangent vectors at* p*. We require that the* e_i *vary in a* C^∞ *manner. For fixed* p*, the basis* $\{e_1, e_2\}$ *is called* the frame at p*. If the surface has a fixed orientation, we say the framing is* oriented *if* e_2 *is obtained from* e_1 *by a counterclockwise rotation.*

Frames can easily be produced. Start with the basis $\frac{\partial}{\partial u}$ and $\frac{\partial}{\partial v}$, and apply the Gram-Schmidt process. This was done on page 40 for the surface $z = y^2 - x^2$, producing the following picture.



Frames on surfaces are the analogues of the moving frame on a curve. But in surface theorem, frames are not unique, so some mathematicians avoid them. We will see that frames are extremely useful.

On the surface of the earth, there is a natural frame at each point except the north and south poles: e_1 points toward the rising run in the east, and e_2 points toward the axis of rotation at the north pole. This frame yields infinitesimal orthonormal coordinates, so inhabitants of the earth learn the Pythagorian theorem in school. But in the large, the frame behaves in an unexpected way due to the curvature of the earth. If occupants follow the e_1's, their journeys are longer than necessary because paths along the e_1's are not geodesics. And if two occupants start one horizontal mile apart and follow e_2's north, they find themselves closer and closer together over time.

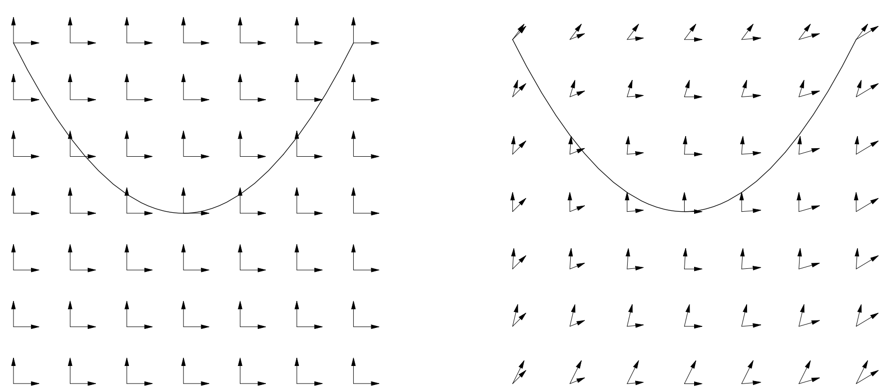
Imagine that the south pole were inhabitable. There is no natural framing there, but it would be inevitable that the government would establish a framing so farms could be divided into rectangular plots and efficient roads could be established.

Remark: Fix a framing. Suppose that $\gamma(s)$ is a path parameterized by arc length. At each time s, $\gamma'(s)$ is a unit vector and thus

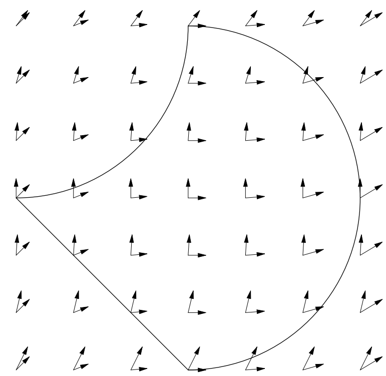
$$\gamma'(s) = \cos\theta(s)\ e_1 + \sin\theta(s)\ e_2$$

for an angle $\theta(s)$ determined up to a multiple of 2π. If we pick a starting angle, there is clearly a unique way to extend $\theta(s)$ to the entire curve.

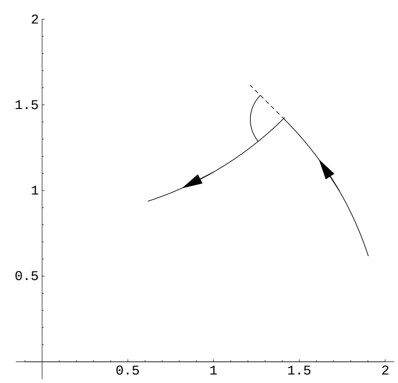
The pictures below show this process in action. The orthonormal basis $\{e_1,\ e_2\}$ may not look orthonormal to us, so we need the g_{ij} to compute $\theta(s)$.



Remark: Suppose that $\mathcal{R}$ is a region in our coordinate system bounded by a finite number of curves meeting at corners. Without loss of generality, we can parameterize each of these boundary curves by arclength, and translate the parameter interval so the first boundary curve $\gamma_1(s)$ is defined on $[0, s_1]$, the second $\gamma_2(s)$ is defined on $[s_1, s_2]$, etc., and the final boundary curve $\gamma_n(s)$ is defined on $[s_{n-1}, s_n]$. Thus the entire curve is defined on $[0, s_n]$. As in the Gauss-Bonnet theorem, we suppose that the boundary curves circle the region counterclockwise, never touching except at the beginning and end.



Choose angles $\theta_1(s)$ along $\gamma_1(s)$ so $\gamma_1'(s) = \cos\theta_1\ e_1 + \sin\theta_1\ e_2$. At s_1, γ_1' makes an angle $\theta_1(s_1)$ with the frame and γ_2' makes an angle $\theta_2(s_1)$ with the frame. Call the exterior angle at this point θ_1. Clearly we can uniquely choose $\theta_2(s)$ so $\theta_2(s) - \theta_1(s)$ equals the exterior angle θ_1.



Continue this process completely around the boundary. The frames allow us to assign tangent angles at each point along the boundary curves. This is a great aid in the proof of the Gauss-Bonnet theorem.

Remark: Define $\Delta\theta_i(s) = \theta_i(s_i) - \theta_i(s_{i-1})$. This number is the amount that the tangent angle of the curve γ_i increases from beginning to end. Notice that this increase is partly due to the curvature of the curve γ_i and partly due to turning of the frame $\{e_1,\ e_2\}$. Disentangling these two causes will occur us in the next few sections. The following theorem shows that, modulo an analysis of $\Delta\theta_i(s)$, we are close to the Gauss-Bonnet theorem:

Theorem 75 *The angle changes $\Delta\theta_i(s)$ and exterior angles θ_i satisfy the equation*

$$\sum_i \theta_i + \sum_i \Delta\theta_i(s) = 2\pi.$$

Proof: The total change in θ must be an integer multiple of 2π because the curve returns to its starting point.

Gradually shrink the boundary curve to a circle. The total change in θ will change continuously under this deformation. Since the total change is always an integer multiple of 2π, it cannot change continuously unless it is constant. So the total change in θ must equal the total change for a small circle. But when we shrink small enough, the frame will become essentially constant, and the theorem reduces to the assertion that the angle increases by 2π as we go around a classical Euclidean circle. QED.

6.11 Differential Forms

To complete the proof of the Gauss-Bonnet theorem, we must prove that

$$\sum \Delta\theta_i(s) = \sum \int_{\gamma_i} \kappa_g + \int\int_{\mathcal{R}} \kappa_1\kappa_2.$$

This will be proved using Green's theorem. The trick is to find an appropriate vector field so the integral of the field over the boundary gives $\sum \int_{\gamma_i} \kappa_g$ and the integral over the interior gives $\sum \int\int_{\mathcal{R}} \kappa_1\kappa_2$. This vector field (which we will call a one-form) will be defined in future sections. First we'd like to review Green's theorem and introduce the notation we will be using.

Definition 27 *A* one-form *ω on a coordinate system is an expression*

$$\omega = \omega_1(u,v)\ du + \omega_2(u,v)\ dv$$

where each ω_i is a C^∞ function on the coordinate domain.

Definition 28 *A* two-form *Ω on a coordinate system is an expression*

$$\Omega = \Omega_{12}(u,v)\ du \wedge dv$$

where Ω_{12} is a C^∞ function on the coordinate domain.

Definition 29 *Let ω be a one-form and let $\gamma(t) = (u(t), v(t))$ be a C^∞ curve, for $a \le t \le b$. The* line integral *of ω over γ is defined to be*

$$\int_\gamma \omega = \int_a^b \left(\omega_1 \frac{du}{dt} + \omega_2 \frac{dv}{dt}\right)\ dt.$$

More explicitly,

$$\int_\gamma \omega = \int_a^b \left(\omega_1(u(t), v(t)) \frac{du}{dt} + \omega_2(u(t), v(t)) \frac{dv}{dt} \right) dt.$$

Definition 30 *Let Ω be a two-form and let $\mathcal{R}$ be a region in the plane. The* surface integral *of Ω over $\mathcal{R}$ is defined to be*

$$\int\int_{\mathcal{R}} \Omega = \int\int_{\mathcal{R}} \Omega_{12}(u, v)\ dudv.$$

Definition 31 *Let ω be a one-form. The* exterior derivative *of ω is the two-form defined by*

$$d\omega = d\left(\omega_1\ du + \omega_2\ dv\right) = \left(\frac{\partial \omega_2}{du} - \frac{\partial \omega_1}{dv}\right) du \wedge dv.$$

Theorem 76 (Green's Theorem) *If γ is a counterclockwise curve bounding a region $\mathcal{R}$ and ω is a one-form on an open set containing γ and $\mathcal{R}$, then*

$$\int_\gamma \omega = \int\int_{\mathcal{R}} d\omega.$$

6.12 The Correct Interpretation

In advanced calculus, you probably computed line integrals of *vector fields*, and wrote Green's theorem using different notation. In that case, the notation and formulas of the previous section may seem puzzling. This optional section explains the new point of view and motivates definitions in the next section.

The line integral of a vector field $E = (E_x, E_y)$ is defined in vector calculus as

$$\int \vec{E} \cdot \frac{d\vec{\gamma}}{dt}\ dt.$$

Green's theorem is then written

$$\int_{\partial\mathcal{R}} \vec{E} \cdot \frac{d\vec{\gamma}}{dt}\ dt = \int\int_{\mathcal{R}} \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right) dxdy.$$

If these theorems were taken as a model, we would expect the one-form ω above to just be a vector field

$$\omega = \Big(E_x,\ E_y\Big).$$

But then the line integral should have been defined using the g_{ij} inner product, which would give a much more complicated formula

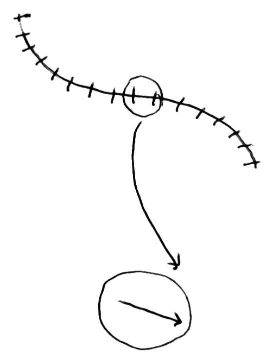
$$\int_a^b \sum_{ij} g_{ij}\ E_i(u(t), v(t))\ \frac{du_j}{dt}\ dt$$

The double integral in our version of Green's theorem is also puzzling, since the integral of a function over the surface always contains an extra factor $\sqrt{g_{11}g_{22} - g_{12}^2}$ by theorem 56. We would expect the double integral in Green's theorem to have the form

$$\int\int \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \right)\ \sqrt{g_{11}g_{22} - g_{12}^2}\ dudv$$

What is going on?

Here is the explanation. Suppose we want to integrate an object ω over curves, but we do not yet know what kind of object ω should be. To integrate, we divide our curve into small pieces, compute ω on each piece, add, and take a limit.



So ω should be an object which gives a number when evaluated on a small piece of the curve. If $\gamma(t)$ is our curve, the small piece from t to $t + \Delta t$ is approximately $\gamma'(t)\ \Delta t$. We expect the world to linearize when we make very small approximations, so

$$\omega(\gamma'(t)\ \Delta t) = \omega(\gamma'(t))\ \Delta t.$$

We conclude that ω should be an object which maps tangent vectors to real numbers. For this reason, the coefficients of our ω should be thought of as entries (ω_1, ω_2) in a matrix defining a transformation from vectors to numbers, and the expression in our line integral is the matrix product

$$\begin{pmatrix} \omega_1 & \omega_2 \end{pmatrix} \begin{pmatrix} \frac{du}{dt} \\ \\ \frac{dv}{dt} \end{pmatrix}$$

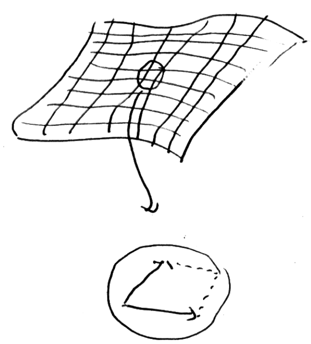
In particular, this expression has no g_{ij}.

In the presence of an inner product $\langle\ ,\ \rangle$, each vector E defines a linear map by $E : X \to \langle E, X\rangle$. In ordinary calculus, the map ω is always assumed to come from a vector E in this way and the theory is written in terms of vector fields.

The same reasoning applies to the double integral in Green's theorem. In this case, the reasoning is easier to understand if we examine the three dimensional version of the theorem, which is called Stokes' theorem.

$$\int_{\partial\mathcal{S}} \vec{E} \cdot \frac{d\vec{\gamma}}{dt}\ dt = \int\int_{\mathcal{S}} \vec{\text{curl}}E \cdot \vec{n}\ d\mathcal{S}.$$

The expression on the right is a surface integral. Suppose we want to integrate an object Ω over a surface, but we do not yet know what kind of object Ω should be. To integrate, we divide the surface into small pieces, compute Ω on each piece, add, and take a limit.



So Ω should be an object which gives a number when evaluated on a small piece of surface. Small parallelograms on the surface are defined by pairs of vectors X and Y. We expect the world to linearize when we make very small approximations, so Ω should be a real valued function defined on pairs of vectors: $\Omega(X, Y)$.

In this case, there is more to be said. Surface integrals depend on an orientation of the surface; changing the orientation changes the sign of the integral. So we expect that $\Omega(Y, X) = -\Omega(X, Y)$. By definition, a two-form is an assignment to each point of space of a map $\Omega(X,Y)$ defined on pairs of vectors, such that $\Omega(Y,X) = -\Omega(X,Y)$. In the special two-dimensional case of interest to us,

$$\Omega\left(X_1\frac{\partial}{\partial u} + X_2\frac{\partial}{\partial v}, Y_1\frac{\partial}{\partial u} + Y_2\frac{\partial}{\partial v}\right) = \Omega\left(\frac{\partial}{\partial u}, \frac{\partial}{\partial v}\right)\ (X_1Y_2 - X_2Y_1)$$

and so our ω has only one coefficient.

In classical three-dimensional calculus, such maps ω also arise from vectors E using the formula

$$\Omega(X,Y) = E\cdot(X\times Y)$$

This explains the appearance of $E\cdot n$ in the surface integral formula. However, in advanced mathematics, Ω more often appears directly as a map on pairs of vectors, and associating each such map with a vector E is more confusing than illuminating.

6.13 Forms Acting on Vectors

Definition 32 *Let ω be a one-form and let X be a tangent vector field. Then $\omega(X)$ is the function defined by*

$$\omega(X) = (\omega_1\ du + \omega_2\ dv)\left(X_1\frac{\partial}{\partial u} + X_2\frac{\partial}{\partial v}\right) = \omega_1X_1 + \omega_2X_2.$$

Definition 33 *Let Ω be a two-form and let X and Y be tangent vector fields. Then $\Omega(X,Y)$ is the function defined by*

$$\Omega(X,Y) = (\Omega_{12}\ du\wedge dv)\left(X_1\frac{\partial}{\partial u} + X_2\frac{\partial}{\partial v},\ Y_1\frac{\partial}{\partial u} + Y_2\frac{\partial}{\partial v}\right) = \Omega_{12}\left(X_1Y_2 - X_2Y_1\right).$$

Remark: If you skipped section 6.9, then these definitions may seem mysterious. We are interested in them for precisely one reason, given by the next theorem. This theorem shows that the curl of ω from classical calculus is closely related to the Lie bracket and directional derivatives we have often used in this course.

Theorem 77 *Let ω be a one-form and suppose X and Y are tangent vector fields. Then*

$$d\omega(X,Y) = X\omega(Y) - Y\omega(X) - \omega([X,Y]).$$

Proof: Notice that $d\omega(X,Y)$ is linear over functions in X and Y, so

$$d\omega\left(X_1\frac{\partial}{\partial u} + X_2\frac{\partial}{\partial v},\ Y_1\frac{\partial}{\partial u} + Y_2\frac{\partial}{\partial v}\right) = \sum_i X_iY_j d\omega\left(\frac{\partial}{\partial u_i}, \frac{\partial}{\partial u_j}\right).$$

We will prove that the expression on the right also has this linearity property. Indeed

$$(fX)\omega(Y) - Y\omega(fX) - \omega([fX,Y]) = fX\omega(Y) - Y(f\omega(X)) - \omega(f[X,Y] - Y(f)X)$$

where we have used the identity $[fX,Y] = f[X,Y] - Y(f)X$ proved on page 127. Using the product rule, the previous equation becomes

$$fX\omega(Y) - Y(f)\omega(X) - fY\omega(X) - f\omega([X,Y]) + Y(f)\omega(X)$$

which simplifies to

$$f\Big(X\omega(Y) - Y\omega(X) - \omega([X,Y)\Big)$$

as desired. The expression $X\omega(Y) - Y\omega(X) - \omega([X,Y)$ changes sign when X and Y are interchanged, so it follows that it is also linear in the second variable.

Since both sides of the equation we wish to establish are linear over functions, we need only check this equation on basis vectors. Both sides are skew symmetric, so it suffices to check the equation when $X = \frac{\partial}{\partial u}$ and $Y = \frac{\partial}{\partial v}$. In this case

$$d\omega(X,Y) = \Omega_{12} = \frac{\partial \omega_2}{du} - \frac{\partial \omega_1}{dv}$$

and, using the fact that $[X,Y] = 0$,

$$X(\omega(Y)) - Y(\omega(X)) - \omega([X,Y]) = \frac{\partial}{\partial u}\omega_2 - \frac{\partial}{\partial v}\omega_1.$$

QED.

6.14 The Moving Frame

In the rest of the proof we never need to refer to the coefficients of the one-forms or two-forms we introduce. So we will call these one forms ω and two forms Ω. Unfortunately, we need to consider several one-forms. These various one-forms will be called ω_{ij}. Notice carefully that the indices determine which form we are using and have nothing to do with the coefficients of these forms.

We have come to the decisive moment in the proof. We want to understand how our frame turns as we travel from point to point. The change in the frame will depend on the direction we move. So we want to study $\nabla_X e_1$ and $\nabla_X e_2$.

In chapter one, the Frenet-Serret formulas were obtained by expressing the derivatives of the canonical frame as a linear combination of vectors in this frame. Similarly in surface theory, it turns out that we should express the derivatives of the e_i in terms of the basis $\{e_1,\ e_2\}$.

Definition 34 *Write*

$$\nabla_X e_j = \sum_i \omega_{ij}(X) e_i.$$

The coefficients of these linear combinations define one-forms w_{ij}, called the connection one-forms *of the surface.*

Remark: The following theorem is an analogue of the Frenet-Serret formulas, and expresses the fact that the frame consists of *orthonormal* vectors.

Theorem 78 $\omega_{ij} = -\omega_{ji}$

Proof: Since the e_i are orthonormal,

$$0 = X(\delta_{ij}) = X \ \langle e_i,\ e_j\rangle = \langle \nabla_X(e_i),\ e_j\rangle + \langle e_i,\ \nabla_X(e_j)\rangle$$

Thus

$$0 = \left\langle \sum_k \omega(X)_{ki} e_k,\ e_j \right\rangle + \left\langle e_i,\ \sum_k \omega(X)_{kj} e_k \right\rangle$$

and so

$$0 = \sum_k \omega(X)_{ki} \langle e_k,\ e_j\rangle + \sum_k \omega_{kj}(X) \langle e_i,\ e_k\rangle = \omega_{ji}(X) + \omega_{ij}(X)$$

QED.

Remark: It follows that $\omega_{11} = \omega_{22} = 0$ and $\omega_{21} = -\omega_{12}$. Consequently, there is only one interesting one-form, ω_{12}. Notice that

$$\nabla_X e_2 = \omega_{12}(X) e_1.$$

A little thought shows that this equation should indeed be true. Since the e_i remain orthornomal, the change of e_2 should be in the e_1 direction. Moveover, if we know how e_2 changes, we can determine how e_1 changed.

Remark: Let $\gamma(s)$ be a curve parameterized by arc length, and let $\theta(s)$ be the angle which the tangent to this curve makes with the frame. Now that we understand how the frame changes from point to poine, we can decompose the change in $\theta(s)$ into two pieces, one describing the curvature of $\gamma(s)$ and one describing the turning of the frame.

Theorem 79 *Suppose $\gamma(s)$ is a curve parameterized by arc length, and $\theta(s)$ is the angle which the tangent to this curve makes with the frame. Let κ_g be the geodesic curvature of this curve. Then*

$$\frac{d\theta}{ds} = \kappa_g + \omega_{12}(\gamma').$$

Proof: We have

$$\gamma'(s) = \cos\theta(s)\ e_1 + \sin\theta(s).$$

Hence

$$\frac{D\gamma'}{ds} = -\sin\theta s \frac{d\theta}{ds}\ e_1 + \cos\theta(s)\frac{d\theta}{ds}\ e_2 + \cos\theta(s)\ \nabla_{\gamma'} e_1 + \sin\theta(s)\ \nabla_{\gamma'} e_2$$

The last two terms can be rewritten

$$\cos\theta(s)\ \omega_{21}(\gamma')\ e_2 + \sin\theta(s)\ \omega_{12}(\gamma')\ e_1.$$

Since $\omega_{21} = -\omega_{12}$, we obtain

$$\frac{D\gamma'}{ds} = \left(\frac{d\theta}{ds} - \omega_{12}(\gamma')\right)\ (-\sin\theta\ e_1 + \cos\theta\ e_2).$$

But $-\sin\theta\ e_1 + \cos\theta\ e_2$ is the vector $\cos\theta\ e_1 + \sin\theta\ e_2$ rotated ninety degrees counter-clockwise, and thus points normal to the curve with the correct orientation. By definition of geodesic curvature, $\frac{D\gamma'}{ds}$ is κ_g times this normal for curves parameterized by arc length. So

$$\kappa_g = \left(\frac{d\theta}{ds} - \omega_{12}(\gamma')\right)$$

and the theorem follows. QED.

6.15 The Structural Equation

This section contains only a single a single equation, known as *Cartan's structural equation.* It is obviously very important.

Theorem 80 *The two-form $d\omega_{12} = \Omega$ has the following properties:*

1. *If e_1, e_2 is any oriented basis,*

$$\Omega(e_1, e_2) = \kappa_1\kappa_2.$$

2. *If $\Omega = \Omega_{12}\ du \wedge dv$, then*

$$\Omega_{12} = \sqrt{g_{11}g_{22} - g_{12}^2}\ \kappa_1\kappa_2.$$

Proof: The definition of w_{12} is:

$$\nabla_Y e_2 = \omega_{12}(Y)\ e_1.$$

Hence

$$\nabla_X \nabla_Y e_2 = X(\omega_{12}(Y))\ e_1 + \omega_{12}(Y)\ \nabla_X(e_1) = X(\omega_{12}(Y))\ e_1 + \omega_{12}(Y)\ \omega_{21}(X)\ (e_2)$$

Since $\omega_{21} = -\omega_{12}$, we conclude that

$$\nabla_X \nabla_Y\ e_2 = X(\omega_{12}(Y))\ e_1 - \omega_{12}(Y)\omega_{12}(X)\ e_2.$$

Subtract the same equation with X and Y interchanged, and subtract the equation for $\nabla_{[X,Y]}\ e_2$, to obtain

$$\left(\nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]}\right)\ e_2 = \left(X(\omega_{12}(Y) - Y\omega_{12}(X) - \omega_{[X,Y]}\right)\ e_1$$

or

$$\left(\nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]}\right)\ e_2 = (d\omega_{12})(X,Y)\ e_1 = \Omega(X,Y)\ e_1.$$

Take the inner product of both sides with e_1 to obtain

$$R(X,Y,e_2,e_1) = \Omega(X,Y).$$

If $X = e_1$ and $Y = e_2$, we have

$$\kappa_1\kappa_2 = -R(e_1,e_2,e_1,e_2) = R(e_1,e_2,e_2,e_1) = \Omega(e_1,e_2).$$

If $X = \frac{\partial}{\partial u} = \alpha e_1 + \beta e_2$ and $Y = \frac{\partial}{\partial v} = \gamma e_1 + \delta e_2$ we have

$$\Omega_{12} = (\Omega_{12}\ du \wedge dv)\left(\frac{\partial}{\partial u}, \frac{\partial}{\partial v}\right) = \Omega(X,Y) = R(X,Y,e_2,e_1)$$

and this equals

$$(\alpha\delta - \beta\gamma)\ R(e_1,e_2,e_2,e_1) = (\alpha\delta - \beta\gamma)\ \kappa_1\kappa_2.$$

But $\sqrt{g_{11}g_{22} - g_{12}^2} = (\alpha\delta - \beta\gamma)$ because

$$g_{11}g_{22} - g_{12}^2 = \langle X,X\rangle\,\langle Y,Y\rangle - \langle X,Y\rangle^2$$

which equals

$$(\alpha^2 + \beta^2)\ (\gamma^2 + \delta^2) - (\alpha\gamma + \beta\delta)^2 = (\alpha\delta - \beta\gamma)^2.$$

QED.

6.16 Proof of Gauss-Bonnet

In section 6.7, we proved that

$$\sum_{\text{exterior angles}} \theta_i + \sum_i \Delta\theta_i(s) = 2\pi.$$

Notice that

$$\Delta\theta_i(s) = \int_{s_{i-1}}^{s_i} \frac{d\theta_i}{ds} = \int_{s_{i-1}}^{s_i} \Big(\kappa_g + \omega_{12}(\gamma'(t))\Big)$$

Since our curve is parameterized by arc length, $\int_{s_{i-1}}^{s_i} \kappa_g$ is just the integral of the function κ_g over the curve. So we have

$$\sum_{\text{exterior angles}} \theta_i \quad + \sum_{\text{boundary curves}} \int_{\gamma_i} \kappa_g \quad + \sum_{\text{boundary curves}} \int_{s_{i-1}}^{s_i} \omega_{12}(\gamma') = 2\pi$$

We claim that $\int_{s_{i-1}}^{s_i} \omega(\gamma')$ is the line integral of ω over γ. Indeed in coordinates $\omega = \omega_1\ du + \omega_2\ dv$ and $\omega(\gamma') = \omega_1 \frac{du}{ds} + \omega_2 \frac{dv}{ds}$ and so

$$\int_{s_{i-1}}^{s_i} \omega(\gamma')\ ds = \int_{s_{i-1}}^{s_i} \left(\omega_1 \frac{du}{ds} + \omega_2 \frac{dv}{ds}\right)\ ds = \int_\gamma \omega.$$

Consequently,

$$\sum_{\text{boundary curves}} \int_{s_{i-1}}^{s_i} \omega_{12}(\gamma') = \int_{\partial\mathcal{R}} \omega_{12}$$

We can apply Green's theorem to convert this to

$$\int\int_{\mathcal{R}} d\omega_{12} = \int\int \Omega = \int\int \Omega_{12}\ du \wedge dv = \int\int_{\mathcal{R}} \sqrt{g_{11}g_{22} - g_{12}^2}\ \kappa_1\kappa_2\ dudv$$

The last integral is the integral of the *function* $\kappa_1\kappa_2$ over the interior of the region, so we obtain

$$\sum_{\text{exterior angles}} \theta_i \quad + \sum_{\text{boundary curves}} \int_{\gamma_i} \kappa_g \quad + \int\int_{\mathcal{R}} \kappa_1\kappa_2 = 2\pi$$

QED.

Chapter 7

Riemann's Counting Argument

7.1 Riemann's Career

In Riemann's day, a candidate for a university position in Germany had to surmount three barriers. The candidate had to write a thesis. Then the candidate had to publish a scholarly paper. Finally, the candidate had to give a probationary lecture for the general public.

Riemann's thesis was on complex variable theory. He introduced *Riemann surfaces*, proved the *Riemann mapping theorem*, and proved the first half of the *Riemann-Roch theorem.* Not bad.

Riemann's scholarly paper was on Fourier series. In this paper, Riemann introduced the *Riemann integral* and used it to investigate convergence of series. OK.

For the probationary lecture, the candidate was asked to provide three topics from which the examining committee would pick one. It was customary for the candidate to list the thesis and the scholarly paper as the first two topics, and customary for the examining committee to pick one of these. Riemann listed his thesis and his scholarly paper, and listed as a third topic *the foundations of geometry.* Riemann's examing committee included Gauss, who convinced the committee to abandon tradition and choose the third topic. "So I am in a quandary," Riemann wrote his father, "since I have to work out this one."

The resulting lecture is one of the most famous mathematical talks in history. Riemann had to proceed without elaborate equations because the lecture was for a general audience. In the lecture, Riemann explained how to generalize Gauss' theory of surfaces to higher dimensions, and gave the following wonderful argument:

7.2 The Counting Argument

In two dimensions, geometry is determined by the fundamental form

$$ds^2 = g_{11}\ dx^2 + 2g_{12}\ dxdy + g_{22}\ dy^2$$

Some choices of g_{ij} give *new* geometries — non-Euclidean geometry, spherical geometry, and geometry on a surface of revolution. Other choices give familiar geometries written in unusual coordinates.

Riemann gave the following counting argument to determine the number of new geometries which could be obtained. The form g_{ij} is determined by three functions. On the other hand, coordinate changes

$$\begin{aligned} r &= \phi(u,v) \\ s &= \psi(u,v) \end{aligned}$$

are determined by two functions. Consequently, we expect the g_{ij} to contain $3 - 2 = 1$ pieces of purely geometric information, determined by one function. *That function, claimed Riemann, is the Gaussian curvature.*

In higher dimensions, Riemann asserted that geometry is determined by a form

$$ds^2 = \sum_{i,j=1}^{n} g_{ij}\ dx_i dx_j$$

Since g_{ij} is symmetric, there are $n + (n-1) + \ldots + 1 = \frac{n(n+1)}{2}$ functions involved. A coordinate change

$$\begin{aligned} y_1 &= \psi_1(x_1, \ldots, x_n) \\ &\cdots \\ y_n &= \psi_n(x_1, \ldots, x_n) \end{aligned}$$

is determined by n functions. Thus there should be $\frac{n(n+1)}{2} - n = \frac{n(n-1)}{2}$ pieces of purely geometric information.

Riemann conjectured that these $\frac{n(n-1)}{2}$ pieces of geometric information could be found as follows. At each point in n-dimensional space, choose a two-dimensional subspace of the tangent space by choosing two basis vectors e_i, e_j from the full basis $e_1, \ldots, e_n$. Follow geodesics out along the plane spanned by e_i and e_j. These geodesics will form a two-dimensional surface in n-dimensions, and this surface will have a Gaussian curvature κ. Nowadays, we call this Gaussian curvature the *sectional curvature of space in the direction spanned by* e_1 *and* e_2.

How many choices do we have for our two-dimensional plane? The vector e_i can be chosen in n ways, and then the vector e_j can be chosen in $n-1$ ways. But the plane spanned by e_i and e_j does not depend on the order of these vectors, so in reality there are $\frac{n(n-1)}{2}$ such choices. *Riemann asserted that the sectional curvatures in these* $\frac{n(n-1)}{2}$ *directions are exactly the* $\frac{n(n-1)}{2}$ *pieces of geometric information hidden in the* g_{ij}.

Traces of this argument can be found in the mathematics of this course. On page 127, we proved that when R is the Riemann curvature tensor and e_1, e_2 is an orthonormal basis, the number $R(e_1, e_2, e_1, e_2)$ is independent of the choice of this basis. In higher dimensions, our proof works without change to prove that whenever e_i and e_j is an orthonormal basis, the number

$$R(e_i, e_j, e_i, e_j)$$

depends only on the two-dimensional space spanned by e_i and e_j, and not on the basis used to compute it. This number equals the sectional curvature up to a sign.

Moreover, it is possible to prove that two candidates for a curvature tensor, $R(X, Y, Z, W)$ and $S(X, Y, Z, W)$, are the same if and only if they give the same sectional curvature for each two-dimensional plane spanned by basis vectors e_i and e_j. So Riemann's information is *exactly* the information hidden in the curvature tensor.

7.3 The Main Theorems

The following theorems provide two cases in which Riemann's insight is verified by formal theorems.

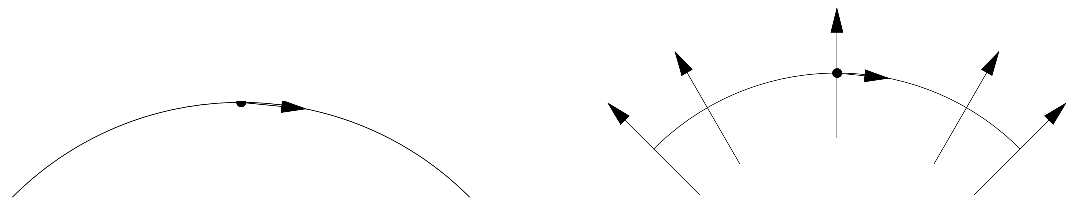
Theorem 81 *Let $\mathcal{S}$ be a surface whose Gaussian curvature is zero. Then it is possible to define local coordinates $s(u, v)$ near any point p so that in these local coordinates, the g_{ij} equal δ_{ij} and*

$$ds^2 = du^2 + dv^2.$$

Theorem 82 *Let $\mathcal{S}$ be a surface whose Gaussian curvature is constant. By magnifying the surface, we may assume that the Gaussian curvature is -1, 0, or 1. Then each point of the surface has a neighborhood isometric to an open set in the sphere, the plane, or non-Euclidean geometry, depending on the value of the curvature.*

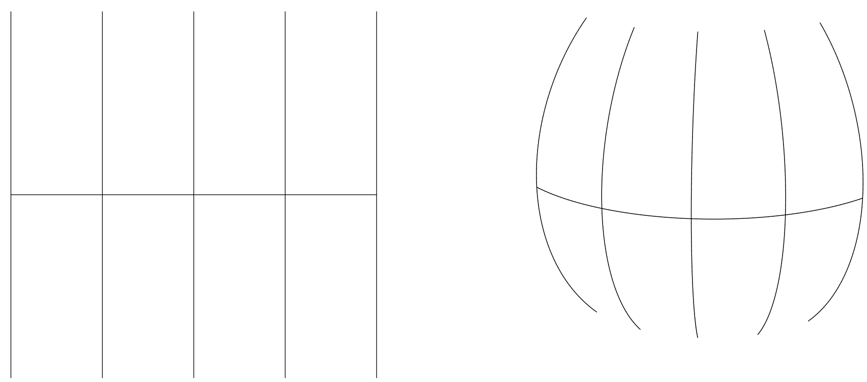
Step 1 of the Proof: Let p be a point in an arbitrary surface $\mathcal{S}$. Choose any geodesic $\gamma(v)$ through p parameterized by arc length and such that $\gamma(0) = p$. Suppose this geodesic is defined for $-\delta < v < \delta$.

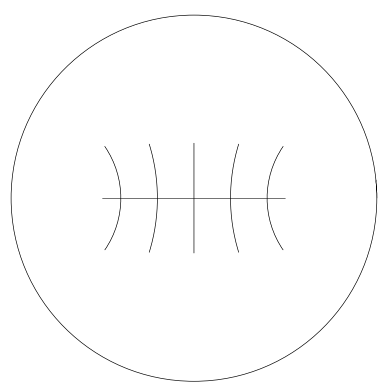
At each point on the geodesic γ, construct the unique geodesic $\tau(u)$ perpendicular to γ. Let τ be parameterized by arc length and suppose that $\tau(0)$ is the starting point on the curve γ. Choose $\epsilon > 0$ so all of these geodesics are defined for $-\epsilon < u < \epsilon$.



Use this construction to produce a coordinate system (u, v) near p. To get to an arbitrary point (u, v), follow γ from p to $\gamma(v)$, and then follow the particular τ which starts at $\gamma(v)$ from $\tau(0)$ to $\tau(u)$.

Below are pictures of this coordinate system in the plane, on the sphere, and on the Poincare disk.





Step 2: If we have two surfaces $\mathcal{S}$ and $\mathcal{T}$ containing points p and q, we can introduce the above coordinate systems in both surfaces, and then map the point (u, v) in $\mathcal{S}$ to the point (u, v) in $\mathcal{T}$. We are going to prove that this map preserves the metric if both surfaces have the same constant curvature. From now on, we work with one surface and one coordinate system. We will discover that we can completely determine the metric g_{ij} in our special coordinate system if the Gaussian curvature is constant. If so, the proof is complete.

Step 3: Let $e_1(u, v) = \frac{\partial s}{\partial u}$ and $e_2(u, v) = \frac{\partial s}{\partial v}$. Then e_1 and e_2 are vector fields in our coordinate system. Since $\tau(u)$ is a geodesic parameterized by arc length, e_1 has length one and $\frac{De_1}{du} = 0$ are every point (u, v). Since $\gamma(v)$ is a geodesic parameterized by arc length, e_2 has length one and $\frac{De_2}{dv} = 0$ at points of the form $(0, v)$. Since τ is perpendicular to γ, $\langle e_1, e_2 \rangle = 0$ at points of the form $(0, v)$.

Step 4: We have

$$\frac{\partial}{\partial u} \langle e_1, e_2 \rangle = \left\langle \frac{De_1}{du}, e_2 \right\rangle + \left\langle e_1, \frac{De_2}{du} \right\rangle$$

But $\frac{De_1}{du} = 0$, so this becomes

$$\frac{\partial}{\partial u} \langle e_1, e_2 \rangle = \left\langle e_1, \frac{De_2}{du} \right\rangle$$

However, according to the fifth item in the main theorem of section 4.4 we have

$$\nabla_{\frac{\partial}{\partial u}} \frac{\partial}{\partial v} - \nabla_{\frac{\partial}{\partial v}} \frac{\partial}{\partial u} = \left[\frac{\partial}{\partial u}, \frac{\partial}{\partial v} \right] = 0$$

and so

$$\frac{De_2}{du} = \frac{De_1}{dv}.$$

Substituting in the second formula of this step, we obtain

$$\frac{\partial}{\partial u} \langle e_1, e_2 \rangle = \left\langle e_1, \frac{De_2}{du} \right\rangle = \left\langle e_1, \frac{De_1}{dv} \right\rangle = \frac{1}{2} \frac{d}{dv} \langle e_1, e_1 \rangle$$

Since e_1 is always a unit vector, this is zero and consequently $\langle e_1, e_2 \rangle$ is independent of u. But $\langle e_1, e_2 \rangle = 0$ on $(0, v)$. Consequently, $\langle e_1, e_2 \rangle$ is always zero.

Step 5: It follows that $g_{11} = 1$ and $g_{12} = 0$. Let $h(u, v) = ||e_2||$. Then $g_{22} = h^2$.

Notice that up to this point, we have made no assumptions about curvature. Therefore, *on any surface, we can introduce new coordinates so the coordinate curves are perpendicular and* $g_{11} = 1, g_{12} = 0, g_{22} = h^2(u, v)$.

Step 6: **Lemma:** We have

1. $h(0, v) = 1$
2. $\frac{\partial h}{\partial u}(0, v) = 0.$

Proof: Since the vector e_2 has length one on $(0, v)$, $h(0, v) = 1$.

Notice that

$$\frac{\partial}{\partial u} h^2 = \frac{\partial}{\partial u} \langle e_2, e_2 \rangle = 2 \left\langle e_2, \frac{De_2}{du} \right\rangle$$

The Lie bracket argument of step 4 converts this into

$$\frac{\partial}{\partial u} h^2 = 2 \left\langle e_2, \frac{De_1}{dv} \right\rangle = 2 \frac{d}{dv} \langle e_2, e_1 \rangle - 2 \left\langle \frac{De_2}{dv}, e_1 \right\rangle = -2 \left\langle \frac{De_2}{dv}, e_1 \right\rangle$$

Since $\gamma(v)$ is a geodesic, $\frac{De_2}{dv} = 0$ on $(0, v)$. We conclude that $\frac{\partial h}{\partial u} = 0$ on $(0, v)$. QED.

Step 7: Since $g_{11} = 1, g_{12} = 0$, and $g_{22} = h^2$, the standard calculations of previous sections rapidly give

$$\Gamma^1_{11} = -h \; \frac{\partial h}{\partial u} \qquad \Gamma^2_{12} = \frac{1}{h} \; \frac{\partial h}{\partial u} \qquad \Gamma^2_{22} = \frac{1}{h} \; \frac{\partial h}{\partial v}$$

By the theorem on the curvature tensor in section 5.4 we have

$$R_{1212} = -h^2 \kappa_1 \kappa_2.$$

We have

$$R_{1212} = \left(\frac{\partial \Gamma^2_{12}}{\partial u} - \frac{\partial \Gamma^2_{11}}{\partial v} + \sum_m \Gamma^m_{12} \Gamma^2_{1m} - \sum_m \Gamma^m_{11} \Gamma^2_{2m} \right) \; h^2$$

and so

$$R_{1212} = \left(\frac{\partial}{\partial u} \left(\frac{1}{h} \; \frac{\partial h}{\partial u} \right) + \Gamma^2_{12} \Gamma^2_{12} \right) \; h^2 = \left(\frac{\partial}{\partial u} \left(\frac{1}{h} \; \frac{\partial h}{\partial u} \right) + \left(\frac{1}{h} \; \frac{\partial h}{\partial u} \right)^2 \right) \; h^2$$

This simplifies to

$$R_{1212} = h \; \frac{\partial^2 h}{\partial u^2} = -h^2 \kappa_1 \kappa_2.$$

We conclude that $\kappa_1 \kappa_2 = \frac{-1}{h} \; \frac{\partial^2 h}{\partial u^2}$. Since we still have made no assumptions about Gaussian curvature, we have proved

Lemma *Any surface can be given a coordinate system in which* $g_{11} = 1$, $g_{12} = 0, g_{22} = h^2(u, v)$ *and in this coordinate system the Gaussian curvature is given by*

$$\frac{-1}{h} \; \frac{\partial^2 h}{\partial u^2}.$$

We conclude that when the curvature is $-1, 0$, or 1, we have

$$\frac{\partial^2 h}{\partial u^2} = h \qquad \frac{\partial^2 h}{\partial u^2} = 0 \qquad \frac{\partial^2 h}{\partial u^2} = -h.$$

Step 8: These differential equations satisfy boundary conditions determined by the lemma in step 6. We conclude that

Lemma: In all cases, we have $g_{11} = 1$ and $g_{12} = 0$.

1. If the Gaussian curvature is minus one, we have $g_{22} = \cosh u$.
2. If the Gaussian curvature is zero, we have $g_{22} = 1$.
3. If the Gaussian curvature is one, we have $g_{22} = \cos u$.

Step 9: Consequently, the g_{ij} are completely determined in our coordinate system by the constant Gaussian curvature. QED.

Corollary 83 *On a surface of constant Gaussian curvature*

1. *if p and q are two points, there is an isometry from a neighborhood of p to a neighborhood of q taking p to q*
2. *if e_1, e_2 and f_1, f_2 are orthonormal bases of tangent vectors at a point p, there is an isometry from a neighborhood of p to a second neighborhood of p, fixing p and taking e_i to f_i.*

7.4 Surfaces of Revolution with Constant Curvature

We end this course by determining those surfaces of revolution which have constant Gaussian curvature.

If a surface is obtained by rotating the function $y = f(x)$ about the x-axis, its Gaussian curvature is

$$\kappa = \frac{-f''}{f(1 + (f')^2)^2}$$

by a calculation on page 96. This will be constant just in case f satisfies the differential equation

$$\frac{d^2 f}{dx^2} = -\kappa \; f(x) \left(1 + \left(\frac{df}{dx} \right)^2 \right)^2$$

In the special case $\kappa = 0$, we have $f(x) = ax + b$ and we obtain cones and cylinders.

Remark: Otherwise we simplify the equation when we parameterize the curve $y = f(x)$ by arc length. The graph of $y = f(x)$ is given as a parameterized curve by $\gamma(x) = (x, f(x))$ and its arc length is

$$s(x) = \int_{x_0}^{x} \sqrt{1 + \left(\frac{df}{dx} \right)^2} \, dx$$

This function can be solved for x in terms of s, giving $x = \varphi(s)$. Let $g(s)$ be the function f written in terms of s instead of x, so that $g(s) = f(x) = f(\varphi(s))$. Notice that $f(x) = g(s(x))$.

Theorem 84 *We have*

$$\frac{d^2 g}{ds^2} = -\kappa g.$$

Consequently when $\kappa = 1$ *we have*

$$g(s) = A \sin(s + \delta)$$

and when $\kappa = -1$ *we have*

$$g(s) = Ae^{s} + Be^{-s}.$$

Proof: By the chain rule, $\frac{df}{dx} = \frac{dg}{ds} \frac{ds}{dx}$ and so

$$\frac{df}{dx} = \frac{dg}{ds} \sqrt{1 + \left(\frac{df}{dx} \right)^2}$$

Differentiating again and using the product and chain rules, we have

$$\frac{d^2 f}{dx^2} = \frac{d^2 g}{ds^2}\left(1+\left(\frac{df}{dx}\right)^2\right) + \frac{dg}{ds}(1/2)\left(1+\left(\frac{df}{dx}\right)^2\right)^{-1/2} 2\frac{df}{dx}\frac{d^2 f}{dx^2}$$

Replace $\frac{df}{dx}$ in the last term with $\frac{dg}{ds}\sqrt{1+\left(\frac{df}{dx}\right)^2}$ to obtain

$$\frac{d^2 f}{dx^2} = \frac{d^2 g}{ds^2}\left(1+\left(\frac{df}{dx}\right)^2\right) + \left(\frac{dg}{ds}\right)^2 \frac{d^2 f}{dx^2}$$

or

$$\frac{d^2 f}{dx^2}\left(1-\left(\frac{dg}{ds}\right)^2\right) = \frac{d^2 g}{ds^2}\left(1+\left(\frac{df}{dx}\right)^2\right)$$

Substitution of the differential equation for f at the start of this section gives

$$-\kappa f(x)\left(1+\left(\frac{df}{dx}\right)^2\right)^2\left(1-\left(\frac{dg}{ds}\right)^2\right) = \frac{d^2 g}{ds^2}\left(1+\left(\frac{df}{dx}\right)^2\right)$$

and this can be rewritten as follows since $f(x) = g(s)$:

$$-\kappa g(s)\left(1+\left(\frac{df}{dx}\right)^2\right)^2\left(1-\left(\frac{dg}{ds}\right)^2\right) = \frac{d^2 g}{ds^2}\left(1+\left(\frac{df}{dx}\right)^2\right)$$

The equation $\frac{df}{dx} = \frac{dg}{ds}\sqrt{1+\left(\frac{df}{dx}\right)^2}$ implies that $1+\left(\frac{df}{dx}\right)^2 = 1+\left(\frac{dg}{ds}\right)^2\left(1+\left(\frac{df}{dx}\right)^2\right)$ and so $\left(1+\left(\frac{df}{dx}\right)^2\right)\left(1-\left(\frac{dg}{ds}\right)^2\right) = 1$. So the previous equation simplifies to

$$-\kappa\ g(s) = \frac{d^2 g}{ds^2}.$$

QED.

Remark: This does not finish the problem. Our curve has the form $\gamma(s) = (x(s), g(s))$ and is parameterized by arc length, so $1 = \left(\frac{dx}{ds}\right)^2 + \left(\frac{dg}{ds}\right)^2$. Therefore $\frac{dx}{ds} = \sqrt{1-\left(\frac{dg}{ds}\right)^2}$ and

$$x(s) = \int \sqrt{1-\left(\frac{dg}{ds}\right)^2}\ ds$$

Theorem 85 *The surfaces of revolution with Gaussian curvature 1 are obtained by rotating $\gamma(s) = (x(s), A\sin(s+\delta))$ around the x-axis, where*

$$x(s) = \pm \int \sqrt{1 - A^2\cos^2(s+\delta)}\, ds$$

The surfaces of revolution with Gaussian curvature -1 are obtained by rotating $\gamma(s) = (x(s), Ae^s + Be^{-s})$ around the x-axis, where

$$x(s) = \pm \int \sqrt{1 - (Ae^s - Be^{-s})^2}$$

Remark: Some of these are elliptic integrals and must be calculated numerically.

Suppose the Gaussian curvature is one. Notice that $A = 1$ gives

$$x(s) = \pm \int \sqrt{1 - \cos^2(s+\delta)} = \pm \int \sin(s+\delta) = \pm\cos(s+\delta)$$

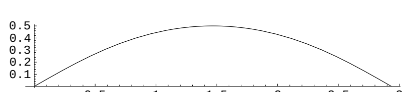
and the surface is obtained by rotating

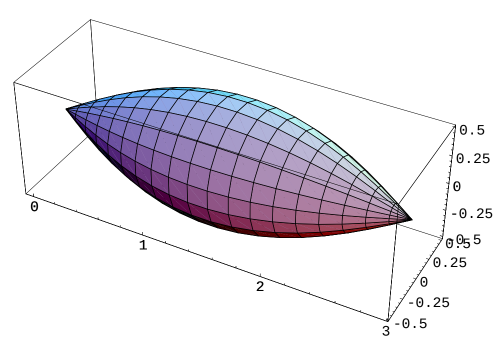
$$\gamma(s) = (\cos(s), \sin(s))$$

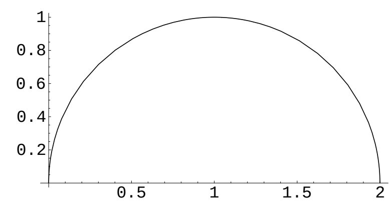
about the x-axis. The resulting surface of revolution is a sphere.

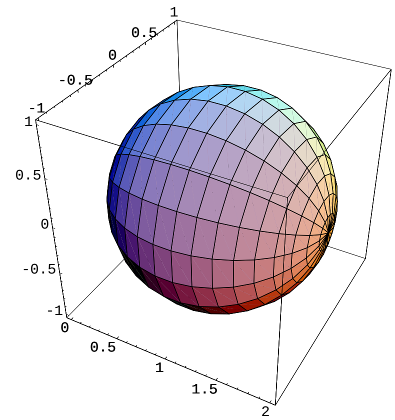
We easily program Mathematica to provide other surfaces of positive curvature. Samples are pictured for $a = 1/2, A = 1$, and $A = 2$. Notice that the remaining surfaces have singularities at the endpoints.

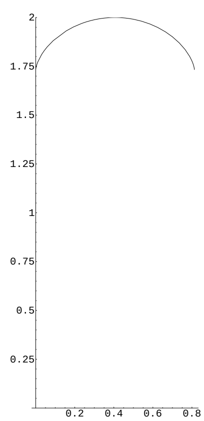
```
x[A_, s0_, s_] := NIntegrate[Sqrt[1 - A^2 Cos[t]^2], {t, s0, s}];

Graph[A_, s0_, bound_] := ParametricPlot[{x[A, s0, s], A Sin[s]},
     {s, s0, bound}, AspectRatio->Automatic];

Surface[A_, s0_, bound_] := ParametricPlot3D[{x[A, s0, s],
     A Sin[s] Cos[theta], A Sin[s] Sin[theta]},
     {s, s0, bound}, {theta, 0, 2 Pi}];
```

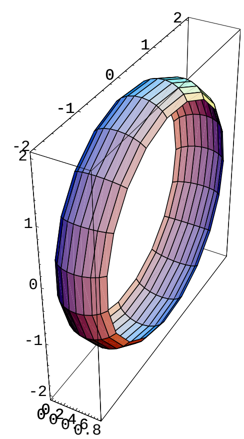












Remark: We will let the reader use Mathematica to provide corresponding surfaces of negative Gaussian curvature. We want to discuss one special case of historical interest. Consider $g(s) = e^s$. Then

$$x(s) = \pm \int_0^s \sqrt{1 - e^{2s}}\ ds$$

The substitution $e^s = \sin\phi$ converts this integral to

$$\int \sqrt{1 - \sin^2\phi}\ \frac{\cos\phi\ d\phi}{\sin\phi} = \int \left(\frac{1}{\sin\phi} - \sin\phi\right)\ d\phi = -\ln(\csc\phi + \text{ctn}\phi) + \cos\phi$$

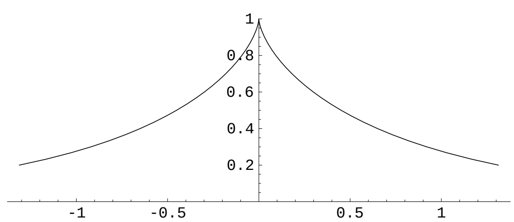
and thus to

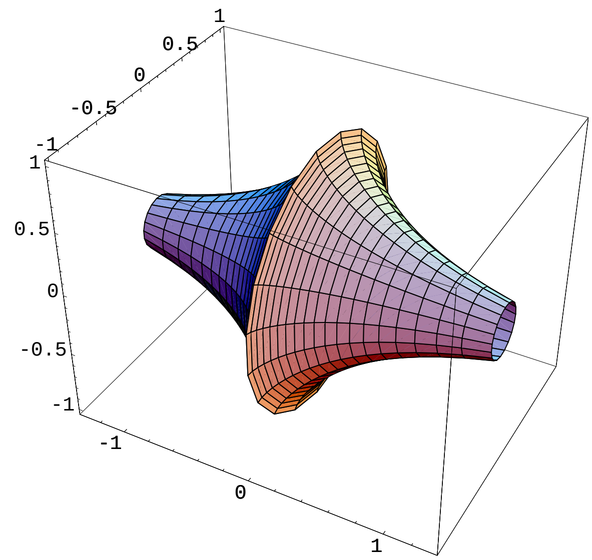
$$-\ln\left(\frac{1 + \sqrt{1 - e^{2s}}}{e^s}\right) + \sqrt{1 - e^{2s}} = s - \ln\left(1 + \sqrt{1 - e^{2s}}\right) + \sqrt{1 - e^{2s}}.$$

Let $e^s = y$. Then our curve is

$$\gamma(y) = \Big(\pm \Big[\ln y - \ln \Big\{ 1 + \sqrt{1 - y^2} \Big\} + \sqrt{1 - y^2} \Big],\ y \Big)$$

The result is a famous surface used often to model non-Euclidean geometry. This surface has a singularity at $x = 0$ and has the wrong topology, so it is only a local model for non-Euclidean geometry. But it was the standard model during most of the 19th century before Poincare's discovery of the disk model, so the surface is often found on the cover of non-Euclidean geometry books. And here is more. And still more.





Bibliography

[1] Albert Einstein. *The Principle of Relativity.* Dover Publications, 1924.

> This inexpensive modern paperbook contains reprints of Einstein's important papers on relativity, and of related papers by other physicists. It includes Lorentz's paper on the Michelson-Morley experiment, Einstein's paper *On the Electrodynamics of Moving Bodies* introducing the special theory, and Einstein's *The Foundation of the General Theory of Relativity* introducing his theory of gravitation.

[2] Euclid. *The Thirteen Books of The Elements.* Dover Publications, Inc., 1956. The three volumes were edited by Sir Thomas L. Heath.

> This wonderful and inexpensive edition has hundreds of pages of notes about the oldest existing manuscripts, the first printed editions, mathematical comments about the postulates over the centuries, and so forth. Volume I contains Euclid's Books 1 and 2. Book 1 is a sort of novel which starts with the construction of an equilaterial triangle with given base and ends with the proof of the Pythagorian theorem and its converse.

[3] Karl F. Gauss. *General Investigations of Curved Surfaces.* Raven Press, 1965.

> An English translation of Gauss' fundamental paper. This edition is probably out of print, but secondhand copies may be available.

[4] Alfred Gray. *Modern Differential Geometry of Curves and Surfaces.* CRC Press, Inc., 1993.

> This book makes heavy use of the computer program *Mathematica*, allowing readers to experiment with the theorems without doing tedious calculations by hand. The book has wonderful illustrations created by the program.

[5] Sigurdur Helgason. *Differential Geometry and Symmetric Spaces.* Academic Press, 1962.

This advanced book has an introductory chapter which covers modern Differential Geometry in a concise manner. All of the fundamental results are proved for arbitrary manifolds in the chapter's 81 pages.

[6] John McCleary. *Geometry from a Differentiable Viewpoint.* Cambridge University Press, 1994.

A wonderful textbook which starts with Euclid's axiomatic approach to geometry and the discovery of Non-Euclidean geometry using Euclid's methods. Then the author introduces the modern theory based on calculus. Ultimately he shows that both approaches are really about the same fundamental ideas.

[7] John Milnor. *Morse Theory*, volume Number 51 of *Annals of Mathematics Studies.* Princeton University Press, 1963.

John Milnor is an important modern mathematician with an elegant writing style. This is his famous book on Morse theory and its connection to algebraic topology. In the later chapters, he needs results from Differential Geometry, and he covers the entire subject in 23 luminous pages, in a section called *A Rapid Course in Riemannian Geometry.* My notes were influenced by Milnor's treatment.

[8] John Milnor. *Topology from a Differentiable Viewpoint.* Princeton Landmarks in Mathematics. Princeton University Press, 1997.

Another famous Milnor book. The main topic is modern differential topology, but Milnor covers the modern definition of a differentiable manifold in the first few pages.

[9] Dirk J. Struik. *Lectures on Classical Differential Geometry.* Addison-Wesley Publishing Company, Inc., 1961.

The standard source for the classic 19th century approach to differential geometry. When I was in college, professors lectured on the modern abstract approach and told students to read Struik for concrete applications. Struik lived a long life and gave a mathematical lecture on his 100th birthday!