Riemannian Geometry

Universität Mannheim

Spring 2024

Ross Ogilvie

Contents

Chapter 1

Curves and Surfaces in $\mathbb{R}^3$

The study of curves and surfaces have a long history in mathematics and have been some of the principal objects of study. In the classical era, we are all familiar with the Euclid's elements with its emphasis on parallel lines, straightedge constructions, triangles, and proportions. These are not very curvy, but perhaps this is because his work on conics (circles, ellipses, parabolas, and hyperbolas) is lost. But surviving works of Archimedes (Measurement of a Circle, On the Sphere and Cylinder, Quadrature of the Parabola) and Apollonius show that curves were a topic of interest and understood in this time.

Though were was some development in the middle ages, particularly in connection with the cubic equation (Khayyam, Viète), and also indirectly in map making (Mercator), general curves were not really considered until the arrival of Cartesian coordinates in the 17th century. A notable early application of this newfound analytical power is seen in the solution to the brachistochrone curve. This problem, posed by Johann Bernoulli in 1696, challenged mathematicians to find the curve along which an object, influenced only by gravity, would travel between two points in the least amount of time. According to his niece, Newton solved the problem literally overnight, using tangents and analysis-type reasoning. In the generations that follow, there was an explosion in the search for finding curves with various special properties.

In the 18th and 19th centuries, the tools of calculus were turned to the study of surfaces, notably with Euler and Gauss exploring different notions of curvatures. A natural question of Lagrange, asking for the surfaces with the least area, had to wait until non-trivial examples could be found which gave hints towards general methods.

But perhaps the most important contribution, and the most important for us in this course, was that of Riemann. All previous mathematicians, as we will do in this first chapter, considered curves and surfaces inside regular old euclidean three dimensional space. Riemann (1868) gave the definition for abstract spaces, which he called manifolds. This definition splits the properties of a space into two type: intrinsic (depending only on the abstract space) and extrinsic (depending how that object is positioned in space). For example, consider two points on a piece of paper. The distance between those points *along the paper* does not depend on whether the piece of paper is laid flat or bent in an arch, whereas the distance *through space* clearly does. The former is an intrinsic property and the latter extrinsic. This program was carried to completion by the 'Italians', who we will meet in later chapters, by around 1900. This was just in time (or

possibly a precondition) for Einstein to use differential geometry in its mature form as a basis the general theory of relativity: our universe is a manifold.

That's enough history; let's see it in action.

1.1 Space Curves and Length

Remark 1.1. In this chapter, and throughout this script unless otherwise stated, we will assume that parameterisations are smooth and injective.

Definition 1.2. *A curve is a smooth function $\alpha : (a, b) \to M$, for some target space M. A path is the restriction of a curve to a closed interval $[\tilde{a}, \tilde{b}] \subset (a, b)$.*

We begin with the example of a helix $\alpha : \mathbb{R} \to \mathbb{R}^3$

$$\alpha(t) = (a\cos t, a\sin t, bt),$$

for constants a, b that describe the size and steepness. In this chapter we will generally consider a parameterised curve $\alpha : [a, b] \to \mathbb{R}^3$, called a *space curve*. Sometimes we will distinguish a parameterised curve α from an un-parameterised curve $\operatorname{img}\alpha$, but other times not. A suitable first question for this curve is to determine its length (or more specifically, the length of any segment of it). In turn, we then must ask how to define the length of a curve. We know how to calculate the distance between points in $\mathbb{R}^3$, so an approximation would be to choose points on the curve, compute the distance between those points, and then add up to total. This approximation will be less than the length, because we are 'cutting corners'. But as we take more and more points into our approximation, it should approach the true value. This leads to

Definition 1.3. *The* length *of a path $\alpha : [a, b] \to \mathbb{R}^3$ is*

$$L(\alpha) = \sup\left\{ \sum_{i=0}^{m-1} \operatorname{dist}(\alpha(t_i), \alpha(t_{i+1})) \mid m \in \mathbb{N}, a = t_0 \le t_1 \le \cdots \le t_m = b \right\}.$$

Let's see if we can use this definition to compute one turn of the helix, $t \in [0, 2\pi]$. The key part of the calculation is the distance between points

$$\begin{aligned}
\operatorname{dist}(\alpha(s), \alpha(t))^2 &= (a\cos s - a\cos t)^2 + (a\sin s - a\sin t)^2 + (bs - bt)^2 \\
&= a^2(2 - 2\cos s\cos t - 2\sin s\sin t) + b^2(s - t)^2 \\
&= a^2(2 - 2\cos(s - t)) + b^2(s - t)^2 \\
&= 4a^2\sin^2\tfrac{1}{2}(s - t) + b^2(s - t)^2.
\end{aligned}$$

This only depends on the difference in parameter values $s - t$. Thus if we break choose the points t_i to be equally space between 0 and 2π in terms of the parameter, each term of the sum will be the same. For this choice we have

$$\sum_{i=0}^{m-1} \operatorname{dist}(\alpha(t_i), \alpha(t_{i+1})) = m \times \sqrt{4a^2\sin^2\tfrac{\pi}{m} + b^2\left(\tfrac{2\pi}{m}\right)^2} = \sqrt{4a^2m^2\sin^2\tfrac{\pi}{m} + 4\pi^2b^2}.$$

Taking the limit as $m \to \infty$ gives $\sqrt{4\pi^2a^2 + 4\pi^2b^2} = 2\pi\sqrt{a^2 + b^2}$.

Exercise 1.4. Complete the proof that this is the length, by showing it is an upper bound.

Though the calculation was straightforward from this example, we should a develop a better method to calculate the length. What we see is that the length will be given by the sum of many small pieces, which should remind you of an integral. Indeed

Theorem 1.5 (Speed)**.** *Let $\alpha : [a, b] \to \mathbb{R}^3$ be a continuously differentiable function. Then*

$$L(\alpha) = \int_a^b \|\alpha'(t)\|\, dt\ .$$

Although it appears we have a simple way to calculate the length of any path, be aware that this integral is often not elementary, with the square root in the norm of the vector being the culprit. This means we must resort to methods to approximate the integral (numerical integration). An example of these difficulties is the length of an ellipse.

Proof. We first show $L(\alpha) \leq \int_a^b \|\alpha'(t)\|\, dt$. Consider therefore a partition $a = t_0 \leq t_1 \leq \ldots \leq t_m = b$ of the interval. We have then the inequality

$$\begin{aligned}\sum_{k=0}^{m-1} d(\alpha(t_k), \alpha(t_{k+1})) = \sum_{k=0}^{m-1} \|\alpha(t_{k+1}) - \alpha(t_k)\| &= \sum_{k=0}^{m-1} \left\| \int_{t_k}^{t_{k+1}} \alpha'(t)\, dt \right\| \\ &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \|\alpha'(t)\|\, dt = \int_a^b \|\alpha'(t)\|\, dt.\end{aligned}$$

Because $L(\alpha)$ is the supremum over all partitions, this gives an upper bound for $L(\alpha)$. This same argument shows that for the restriction of the curve $\alpha|_{[t_1,t_2]}$ (with $a \leq t_1 < t_2 \leq b$) we have

$$(*) \qquad L(\alpha|_{[t_1,t_2]}) \leq \int_{t_1}^{t_2} \|\alpha'(t)\|\, dt.$$

Consider now the following two functions

$$\begin{aligned} s &: [a, b] \to \mathbb{R},\ t \mapsto L(\alpha|_{[a,t]}) \\ \widetilde{s} &: [a, b] \to \mathbb{R},\ t \mapsto \int_a^t \|\alpha'(u)\|\, du\ .\end{aligned}$$

These are meant to capture the length at the parameter t from the start of the path, measured in two ways. Our strategy to finish the proof is not to show the reverse inequality directly. Rather we will show that these two functions are equal. Clearly they are equal at $t = a$.

Observe that s has the property that $L(\alpha|_{[t_1,t_2]}) = s(t_2) - s(t_1)$, and likewise $\int_a^t \|\alpha'(u)\|\, du = \widetilde{s}(t_2) - \widetilde{s}(t_1)$. Inequality (*) above then says $s(t_2) - s(t_1) \leq \widetilde{s}(t_2) - \widetilde{s}(t_1)$. It then follows that

$$\|\alpha(t_2) - \alpha(t_1)\| = d(\alpha(t_1), \alpha(t_2)) \leq L(\alpha|[t_1, t_2]) = s(t_2) - s(t_1) \leq \widetilde{s}(t_2) - \widetilde{s}(t_1).$$

After dividing by $t_2 - t_1$, we arrive at

$$\left\| \frac{\alpha(t_2) - \alpha(t_1)}{t_2 - t_1} \right\| \leq \frac{s(t_2) - s(t_1)}{t_2 - t_1} \leq \frac{\widetilde{s}(t_2) - \widetilde{s}(t_1)}{t_2 - t_1}.$$

When we take the limit as $t_2 \to t_1$, the left and right terms both tend to $\|\alpha'(t_1)\|$, by the definition of derivative and the fundamental theorem of calculus respectively. By the squeeze law $s'(t) = \widetilde{s}'(t)$ and therefore $s(t) = \widetilde{s}(t)$. □

The function s in the above proof (using either definition) is called the *arc-length* function. We see that its derivative is $\|\alpha'\|$, which we call the speed of the parameterisation. Clearly s is weakly monotonically increasing, since it it the integral of a non-negative function. If it is strongly monotonically increasing, then it is a bijective function from $[a, b]$ to $[0, L(\alpha)]$. In this case we can use it to give a new parameterisation of the same path. The advantage of this new parameterisation is that to find the length between two points, we can just subtract their parameter values. For obvious reasons this is called the arc-length parameterisation of a curve.

Example 1.6. In the case of the helix, the speed is

$$\begin{aligned} \alpha'(t) &= (-a\sin t, a\cos t, b) \\ \|\alpha'(t)\| &= \sqrt{a^2 + b^2}, \end{aligned}$$

which is constant. The arc-length function is simply $s(t) = t\sqrt{a^2 + b^2}$, and the inverse is $t(s) = s/\sqrt{a^2 + b^2}$. Hence the helix with arc-length parameterisation is

$$\alpha(s) = \left(a\cos\frac{s}{\sqrt{a^2+b^2}}, a\sin\frac{s}{\sqrt{a^2+b^2}}, \frac{bs}{\sqrt{a^2+b^2}} \right).$$

Note that it is common practice to reuse the name of the curve, even though strictly speaking it is a new function, $\tilde{\alpha} = \alpha \circ t(s)$.

The arc-length is an intrinsic property of the curve. One can imagine an ant crawling along a piece of string, counting the distance as it goes. In fact this is the only intrinsic invariant for a curve, though we do not yet have a clear definition of intrinsic, so we cannot yet prove this.

A sufficient condition that there exists an arc-length parameterisation is that $\|\alpha'(t)\| \neq 0$ for any t. We say that a curve with such a parameterisation is *regularly parameterised*. This condition also serves to rule out some other curves that have undesirable properties.

Example 1.7. Consider $\beta(t) = (t^2, t^3)$. This is called the cusp curve. We have that $\|\beta'\|^2 = 4t^2 + 9t^3$, and in particular vanishes for $t = 0$. If you look at a plot of this curve, you see that it has a cusp singularity at the origin. This is an example of a curve we would like to avoid.

1.2 Osculating Circles

We have already identified $\|\alpha'\|$ with the speed of the parameterisation, but what is α' itself? Naturally it is the tangent vector and it spans the tangent line to the curve. Recall the tangent line is the limit of the line that passes through the points $\alpha(t)$ and $\alpha(t+h)$ as $h \to 0$. Let us consider the generalisation of this to three points $\alpha(t), \alpha(t+h), \alpha(t-h)$ as $h \to 0$. In $\mathbb{R}^3$, three points span a plane, so long as they do not happen to all lie on the same line. The plane which is the limit of these planes is called the *osculating plane* of the curve.

The key to describing the osculating plane is to find two linearly independent vectors that lie in it. Clearly the tangent vector lies in it. The vectors $\alpha(t+h) - \alpha(t)$ and $\alpha(t-h) - \alpha(t)$ lie in the plane, hence their sum does too. We compute

$$\lim_{h\to 0} \frac{\alpha(t+h) - 2\alpha(t) + \alpha(t-h)}{h^2} = \lim_{h\to 0} \frac{\frac{\alpha(t+h)-\alpha(t)}{h} - \frac{\alpha(t)-\alpha(t-h)}{h}}{h} = \alpha''(t).$$

Thus the osculating plane is spanned by the first and second derivative. Moreover, if we use the arc-length parameterisation the we know that the tangent vector always has length 1. If we differentiate the equation $\alpha' \cdot \alpha' = 1$ then we get $\alpha'' \cdot \alpha' = 0$. In words, in this parameterisation the first and second derivatives are an orthogonal basis of the osculating plane.

In fact we can extract even more information from this three point construction. Three points determine not just a plane, but a circle within that plane. The limit of this circle as these three points come together is called the *osculating circle*. We think of it as the 'tangent circle', just like we have a tangent line.

How should we calculate the osculating circle? As for any circle, we should find its center and radius. Conceptually we can find the center by considering the chords $\alpha(t+h) - \alpha(t)$ and $\alpha(t-h) - \alpha(t)$, taking their perpendicular bisectors, and finding the intersection point. Practically, the difficulty is writing down the perpendicular bisector. Let's put this difficulty aside for a moment, and suppose that we have an operator R_h that rotates by a right angle the plane in spanned by the three points with origin $\alpha(t)$. Then the center is the intersection point:

$$\begin{aligned} c &= \alpha(t) + \frac{1}{2}(\alpha(t+h) - \alpha(t)) + uh^{-1}R_h(\alpha(t+h) - \alpha(t)) \\ &= \alpha(t) + \frac{1}{2}(\alpha(t-h) - \alpha(t)) - vh^{-1}R_h(\alpha(t-h) - \alpha(t)). \end{aligned}$$

This is a vector equation in a plane, so u and v are determined by this equation. Really c, u and v are functions of h, since for every h we have a different plane. Taking the limit $h \to 0$ we obtain $c(0) = \alpha(t) + u(0)R_0\alpha'(t)$, so the radius of the osculating circle is $u(0)\|\alpha'(t)\|$. It remains to find $u(0)$.

Rearranging the intersection equation gives

$$\frac{\alpha(t+h) - \alpha(t-h)}{2} = -vh^{-1}R_h(\alpha(t-h) - \alpha(t)) - uh^{-1}R_h(\alpha(t+h) - \alpha(t)).$$

If we just take the limit $h \to 0$ we see that

$$0 = -v(0)R_0(-\alpha'(t)) - u(0)R_0(\alpha'(t)).$$

Since $\alpha'(t) \neq 0$ it must be that $u(0) = v(0)$ (this justifies the choice of sign and the h^{-1} in the equations of the bisectors). If we instead first divide by h and then take the limit, we obtain

$$\begin{aligned}
\alpha'(t) &= -\lim_{h\to 0} vR_h \frac{\alpha(t-h)-\alpha(t)}{h^2} + uR_h \frac{\alpha(t+h)-\alpha(t)}{h^2} \\
&= -\lim_{h\to 0} (v-u)R_h \frac{\alpha(t-h)-\alpha(t)}{h^2} + uR_h \frac{\alpha(t-h)-\alpha(t)+\alpha(t+h)-\alpha(t)}{h^2} \\
&= -\lim_{h\to 0} \frac{v-u}{h} R_h \frac{\alpha(t-h)-\alpha(t)}{h} + uR_h \frac{\alpha(t-h)-2\alpha(t)+\alpha(t+h)}{h^2} \\
&= -(v'(0)-u'(0))R_0(-\alpha'(t)) + u(0)R_0\alpha''(t).
\end{aligned}$$

Dot product both sides with $\alpha'(t)$:

$$\begin{aligned}
\alpha'(t)\cdot\alpha'(t) &= u(0)\alpha'(t)\cdot R_0\alpha''(t) \\
u(0) &= \frac{\alpha'(t)\cdot\alpha'(t)}{\alpha'(t)\cdot R_0\alpha''(t)}
\end{aligned}$$

Thus we have determined $u(0)$, up to the rotation operator R_0. If the curve is parameterised by arc-length, then $\|\alpha'\| = 1$, so the radius of the osculating circle is $u(0)$. Moreover α'' is perpendicular to α', so $\alpha'(t)\cdot R_0\alpha''(t)$ is just the length of α''. In summary we have proved

Theorem 1.8 (Curvature)**.** *For a regular arc-length parameterised curve α, the radius of the osculating is κ^{-1}, where*

$$\kappa(s) = \left\| \frac{d^2\alpha}{ds^2} \right\|,$$

a quantity called the curvature*. The radius of the osculating circle is also called the* radius of curvature*.*

Example 1.9. Let's apply this to the helix. We use the arc-length parameterisation

$$\begin{aligned}
\alpha(s) &= \left(a\cos\frac{s}{\sqrt{a^2+b^2}}, a\sin\frac{s}{\sqrt{a^2+b^2}}, \frac{bs}{\sqrt{a^2+b^2}} \right) \\
\alpha'(s) &= \frac{1}{\sqrt{a^2+b^2}} \left(-a\sin\frac{s}{\sqrt{a^2+b^2}}, a\cos\frac{s}{\sqrt{a^2+b^2}}, b \right) \\
\alpha''(s) &= -\frac{a}{a^2+b^2} \left(\cos\frac{s}{\sqrt{a^2+b^2}}, \sin\frac{s}{\sqrt{a^2+b^2}}, 0 \right) \\
\kappa(s) &= \frac{a}{a^2+b^2}.
\end{aligned}$$

Consider the special case $b = 0$, then our helix is a circle and the curvature is a^{-1}. Increasing a decreases the curvature and vice versa. This matches our intuition, a car driving on a large circle only needs to turn slowly. For the general case, we see that increasing either a or b decreases curvature. We also see that α'' points towards the central axis of the helix.

Exercise 1.10. Notice in the above argument that the rotation operator R_h is defined as a rotation of the plane spanned by the three points and only applied to vectors that lie that plane. We could extend R_h to a linear operator on $\mathbb{R}^3$ by declaring that it preserves vectors perpendicular to the plane. Then the operator can be applied to any vector. Simplify the above calculation by using this observation and the second order Taylor polynomial $\alpha(t+h) = \alpha(t) + h\alpha'(t) + \frac{1}{2}h^2\alpha''(t) + O(h^3)$.

Exercise 1.11. Suppose that α is a regular curve but do not assume that it is parameterised by arc-length. Show that the curvature can be calculated with the formula

$$\kappa(t) = \frac{\|\alpha'(t) \times \alpha''(t)\|}{\|\alpha'(t)\|^3}.$$

1.3 Frenet-Serret Equations

Let us return to our thought experiment of an ant on a piece of string. The ant cannot see the curvature of a piece of string. If the string was a helix, but then you straighten it out into a line, none of the distances along the string have changes. Therefore it must be that curvature is an extrinsic property of the curve. But the curvature is invariant under proper euclidean motions (translations and rotations). This is easy to prove, if $\beta(s) = O\alpha(s) + b$ where O is a rotation and b is a vector, then $\|\beta'(s)\| = \|O\alpha'(s)\| = 1$ shows that β is also parameterised by arc-length and $\beta''(s) = O\alpha''(s)$ shows that the curvatures are equal, because rotation does not change the length of a vector.

It turns out for curves in $\mathbb{R}^3$ up to proper euclidean motions there are only two extrinsic properties: curvature and torsion. The goal of this section is to find the torsion and prove that there are no other extrinsic invariants.

We have seen that for an arc-length parameterised curve, the first and second derivatives form an orthogonal basis of the osculating plane. It is customary to normalise them to an orthonormal basis. $T(s) := \alpha'(s)$ is already unit length and we set $N(s) := \kappa(s)^{-1}\alpha''(s)$. They are called the unit tangent and unit normal vectors respectively. Additionally we define $B(s) = T(s) \times N(s)$, called the unit binormal vector, so that we have an orthonormal basis of $\mathbb{R}^3$. This basis is also known as the Frenet frame. We can also recover a curve given knowledge of this basis by integrating $T(s)$, up to a translation

$$\alpha(s) = \int_a^s T(u)\,du + \alpha(a).$$

The advantage of using a basis that comes from the curve, is that if the curve is rotated, this basis is rotated too. If we use this basis to measure the curve, then we are using the curve to measure itself. We have seen this already in the fact that the curvature can be computed as the length of the second derivative. Let us then investigate the derivatives of the other basis vectors. All these vectors are unit length, so they are perpendicular to their derivatives, eg $N \cdot N = 1$ implies $2N \cdot N' = 0$. Further

$$\begin{aligned} 0 = T \cdot N \quad &\Rightarrow \quad 0 = T' \cdot N + T \cdot N' = \kappa + T \cdot N' \\ 0 = T \cdot B \quad &\Rightarrow \quad 0 = T' \cdot B + T \cdot B' = 0 + T \cdot B' \\ 0 = N \cdot B \quad &\Rightarrow \quad 0 = N' \cdot B + N \cdot B'. \end{aligned}$$

From the second equation, we see that B' is perpendicular to T as well as B. Therefore it is a scalar of N.

Definition 1.12. *The torsion τ of a curve is defined by the formula $B'(s) = -\tau(s)N(s)$. The minus sign is the most common convention, but be aware some authors choose the opposite sign.*

The binormal B is perpendicular to the osculating plane and the tangent vector is always in it. Therefore the torsion tells us how quickly the osculating plane is rotating around the tangent vector. For a curve that lies entirely in a plane, the torsion is zero.

Example 1.13. For the helix, we already know

$$T(s) = \frac{1}{\sqrt{a^2+b^2}}\left(-a\sin\frac{s}{\sqrt{a^2+b^2}}, a\cos\frac{s}{\sqrt{a^2+b^2}}, b\right)$$
$$N(s) = -\left(\cos\frac{s}{\sqrt{a^2+b^2}}, \sin\frac{s}{\sqrt{a^2+b^2}}, 0\right).$$

Therefore

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

We should now differentiate this

$$B'(s) = \frac{b}{a^2+b^2}\left(\cos\frac{s}{\sqrt{a^2+b^2}}, \sin\frac{s}{\sqrt{a^2+b^2}}, 0\right).$$

As expected, this is a scalar multiple of the normal vector. We can read off the torsion

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

Helices with $b > 0$ and $b < 0$ are mirror images of one another. Ones that have $b > 0$ are called right-handed, in accordance with the 'right hand rule'. Here we see that our sign convention of torsion gives right-handed helices positive torsion and left-handed helices negative torsion.

We were not yet done with the derivatives of the basis vectors. From the third equation, we have $N' \cdot B = -N \cdot (-\tau N) = \tau$ and from the first $N' \cdot T = -\kappa$. Thus $N'(s) = -\kappa T + \tau B$. Together these derivatives are called the Frenet-Serret formulas

$$\frac{d}{ds}\begin{pmatrix} T \\ N \\ B \end{pmatrix} = \begin{pmatrix} \kappa N \\ -\kappa T + \tau B \\ -\tau N \end{pmatrix} = \begin{pmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{pmatrix}\begin{pmatrix} T \\ N \\ B \end{pmatrix}$$

Theorem 1.14 (Local uniqueness of curves)**.** *Let α, β be two smooth regular arc-length parameterised curves in $\mathbb{R}^3$ with the same curvature and torsion. For simplicity assume that the curvature is always positive. Then there is a proper euclidean motion (translation and rotation) that takes one to the other.*

Proof. Both bases $\{T_\alpha(0), N_\alpha(0), B_\alpha(0)\}$ and $\{T_\beta(0), N_\beta(0), B_\beta(0)\}$ are right-handed orthonormal bases of $\mathbb{R}^3$. Therefore there is a rotation O that transforms one into the other. Let $b = \beta(0) - O\alpha(0)$. Define $\gamma(s) = O\alpha(s) + b$. It is also arc-length parameterised and $\gamma(0) = O\alpha(0) + \beta(0) - O\alpha(0) = \beta(0)$. Moreover, it has the same curvature and torsion as α and β. By differentiating, we have $T_\gamma(0) = OT_\alpha(0) = T_\beta(0)$ and $\kappa_\gamma N_\gamma = \kappa_\alpha O N_\alpha$, which implies $N_\gamma(0) = ON_\alpha(0) = N_\beta$. By the definition of binormals, $B_\gamma(0) = B_\beta(0)$.

It remains to show that $\beta(s) = \gamma(s)$ for all s. We give two proofs of this fact. The first lies in the observation that the Frenet-Serret formulas are in fact a nine-dimensional system of ODEs (three coordinates for each of the three vectors). By Picard-Lindelöff we know that the initial value problem, which both β and γ satisfy, has a unique solution. In particular $T_\beta(s) = T_\gamma(s)$ for all s. But we can integrate this to see that $\beta(s) = \gamma(s) + c$ for some constant c. Evaluation at $s = 0$ shows that $c = 0$.

We can also use apply the Frenet-Serret formulas directly to show uniqueness

$$\begin{aligned}
&\frac{1}{2}\frac{d}{ds}\left(\|T_\beta - T_\gamma\|^2 + \|N_\beta - N_\gamma\|^2 + \|B_\beta - B_\gamma\|^2\right) \\
&\quad = (T_\beta - T_\gamma)\cdot(T'_\beta - T'_\gamma) + (N_\beta - N_\gamma)\cdot(N'_\beta - N'_\gamma) + (B_\beta - B_\gamma)\cdot(B'_\beta - B'_\gamma) \\
&\quad = (T_\beta - T_\gamma)\cdot\kappa(N_\beta - N_\gamma) + (N_\beta - N_\gamma)\cdot\Big(-\kappa(T_\beta - T_\gamma) + \tau(B_\beta - B_\gamma)\Big) \\
&\qquad\quad + (B_\beta - B_\gamma)\cdot(-\tau)(N_\beta - N_\gamma) \\
&\quad = 0.
\end{aligned}$$

Therefore the sum of squares of the differences is constant. Because it is zero at $s = 0$, it must stay zero for all s. In particular, the tangent vectors are equal. The argument can be finished similarly to the other proof. □

This shows that a curve is uniquely determined in $\mathbb{R}^3$ up to proper euclidean motion by κ and τ. This proves our assertion that these are the only two invariants. Thus the only curves whose curvature and torsion are constant functions are helices. As a special case, the circle is the only curve in the plane with constant curvature (torsion is zero).

In the first half of this chapter dealing with curves, we have already seen several important themes. First there is the difference between intrinsic and extrinsic quantities. Second is the use of special coordinates, in this case arc-length parameterisations. And finally is the idea of measuring the change of vector fields to learn about a space. All three of these ideas will occur repeatedly throughout this course.

Exercise 1.15. Show for a helix

$$a = \frac{\kappa}{\kappa^2 + \tau^2}, \qquad b = \frac{\tau}{\kappa^2 + \tau^2}.$$

1.4 Surfaces

In the second half of this chapter, we consider the example of the helicoid $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$

$$\Phi(u, v) = (u \cos v, u \sin v, bv).$$

We see that for any fixed value of u that we have a helix. Conversely, for any fixed value of v we have a straight line in a plane with $z = bv$. Surfaces which are 'made up of' straight lines are called *ruled surfaces* and were a major topic of study in the theory of surfaces, though we will not go too far down that path. We will also use Φ to represent a general parameterised surface from an open subset $U \subset \mathbb{R}^2$ in $\mathbb{R}^3$. Again, we blur the distinction between a parameterised surface Φ and an un-parameterised surface $\Sigma = \operatorname{img} \Phi$.

We will try to reproduce many of the useful tools for curves in the situation of surfaces. Given a curve $\tilde{\alpha}$ in U, that is a function $\tilde{\alpha} : (a, b) \to U$, we can compose it with Φ to get a curve on $\mathbb{R}^3$ on the surface. We will call this curve $\alpha = \Phi \circ \tilde{\alpha}$ a space curve and $\tilde{\alpha}$ a curve in coordinates if it necessary to distinguish them. The helices and straight lines of the previous sections are then $\tilde{\alpha}_v(t) = (t, v)$ and $\tilde{\beta}_u(t) = (u, t)$. The tangent vector of a general curve $\Phi \circ \tilde{\alpha}$ on the surface can be computed using the chain rule

$$\frac{d}{dt}(\Phi \circ \tilde{\alpha}) = \frac{\partial \Phi}{\partial u}\frac{d\tilde{\alpha}^1}{dt} + \frac{\partial \Phi}{\partial v}\frac{d\tilde{\alpha}^2}{dt} = \begin{pmatrix} \frac{\partial \Phi}{\partial u} & | & \frac{\partial \Phi}{\partial v} \end{pmatrix} \tilde{\alpha}'.$$

In this way we see that $\frac{\partial \Phi}{\partial u}$ and $\frac{\partial \Phi}{\partial v}$ are a basis for the tangent vectors to the surface, called the coordinate basis vectors.

There are two considerations to make now. Just as we restrict ourselves to regular curves, so too should we restrict ourselves to regular surfaces.

Definition 1.16. *A parameterised surface $\Phi : U \to \mathbb{R}^3$ is called* regular *if $J\Phi$ (the Jacobian of Φ, the matrix of partial derivatives) has rank two at every point. Equivalently, if the vectors $\frac{\partial \Phi}{\partial u}$ and $\frac{\partial \Phi}{\partial v}$ are linearly independent.*

The relation to regular curves should be clear. If the two vectors are linearly dependent at some point $\Phi(c^1, c^2)$, i.e. $w^1 \frac{\partial \Phi}{\partial u} + w^2 \frac{\partial \Phi}{\partial v} = 0$, then the curve $\tilde{\alpha}(t) = (w^1 t + c^1, w^2 t + c^2)$ would produce a curve α on the surface that was not regular.

The second consideration would be to try to make an 'arc-length' parameterisation. However this is not possible for a surface, for deep reasons that we will explore in the chapter on curvature. For now we can gain a simple understanding through a thought experiment on the sphere. Suppose that there existed a parameterisation like longitude-latitude coordinates on the unit sphere, but instead of angle it used distance. Choose a point on the equator. We can walk the distance π east along a latitude. This is half-way around the sphere. Then we can go a short distance ε north, walk distance π west, and then ε south. In the coordinate chart, this is a rectangle and we are back where we started. But on the sphere, the line of latitude north of the equator is shorter than the equator, so when we walked a distance π on it, we walked too far. At the end we ended up west of our starting point. You have probably already experienced this problem, because no map of the earth can represent the distances correctly to scale. Some maps represent the distances accurately in one direction (cylindrical equidistant projection preserves distance

on lines of longitudes, azimuthal equidistant projection preserves distance on radial lines), but most maps do not try to represent distance at all and instead try to preserve angle or area.

So if we can't find a parameterisation of a surface that is arc-length in every direction, what data do we need to be able to calculate angle and distance in a given parameterisation? We know from Theorem 1.5 that to calculate length of a curve we only need the length of the tangent vectors. In $\mathbb{R}^3$ both the length of vectors and the angle between two vectors is given by the dot product

$$\|v\| = \sqrt{v \cdot v}, \qquad \operatorname{ang}(v, w) = \frac{v \cdot w}{\|v\| \, \|w\|}.$$

Abstractly the dot product is an example of an inner product: a function on pairs of vectors that is bilinear, symmetric, and positive definite. The restriction of an inner product to a subspace is again an inner product. Thus we can restrict the inner product of $\mathbb{R}^3$ (the dot product) to an inner product on the tangent space of the surface at any point. This is called the *first fundamental form*. It has the symbol I or g. The word 'form' is an old fashioned term for a function from vectors to scalars, which still appears in certain names.

Practically, how can we describe g at some point? We know that every tangent vector to the surface is in the span of $\frac{\partial \Phi}{\partial u}$ and $\frac{\partial \Phi}{\partial v}$. So we compute

$$\begin{aligned}
&g\left(v^1 \tfrac{\partial \Phi}{\partial u} + v^2 \tfrac{\partial \Phi}{\partial v}, w^1 \tfrac{\partial \Phi}{\partial u} + w^2 \tfrac{\partial \Phi}{\partial v}\right) \\
&= v^1 w^1 g\left(\tfrac{\partial \Phi}{\partial u}, \tfrac{\partial \Phi}{\partial u}\right) + v^1 w^2 g\left(\tfrac{\partial \Phi}{\partial u}, \tfrac{\partial \Phi}{\partial v}\right) + v^2 w^1 g\left(\tfrac{\partial \Phi}{\partial v}, \tfrac{\partial \Phi}{\partial u}\right) + v^2 w^2 g\left(\tfrac{\partial \Phi}{\partial v}, \tfrac{\partial \Phi}{\partial v}\right) \\
&= \begin{pmatrix} v^1 & v^2 \end{pmatrix} \begin{pmatrix} \frac{\partial \Phi}{\partial u} \cdot \frac{\partial \Phi}{\partial u} & \frac{\partial \Phi}{\partial u} \cdot \frac{\partial \Phi}{\partial v} \\ \frac{\partial \Phi}{\partial v} \cdot \frac{\partial \Phi}{\partial u} & \frac{\partial \Phi}{\partial v} \cdot \frac{\partial \Phi}{\partial v} \end{pmatrix} \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \\
&=: v^T \begin{pmatrix} g_{1\,1} & g_{1\,2} \\ g_{2\,1} & g_{2\,2} \end{pmatrix} w.
\end{aligned}$$

This is called writing g with respect to the coordinate basis. The symmetry of the dot product means that $g_{i\,j}$ is a symmetric matrix, $g_{1\,2} = g_{2\,1}$. The point here is not that we have some short-cut to avoid taking dot products; when you are doing an example it will often be fast to use the dot product in $\mathbb{R}^3$. The point is that we only need part of the information of the dot product of $\mathbb{R}^3$, namely how it acts on the tangent space of the surface, to compute lengths and angles on the surface. We are trying to separate intrinsic information from extrinsic information. If we allow the point on the surface to vary, then we obtain functions $g_{i\,j}(u, v) : U \to \mathbb{R}$.

Remark 1.17. Observe here that v^i and w^i have superscripts instead of subscripts: these are not powers! This is part of a larger notational convention in the field. It's a little annoying at first, especially when you have to write $(v^2)^2$, but it's worth it in the long run. Roughly speaking, coordinates and components of vectors should use superscripts, and forms should use subscripts.

Example 1.18. For the helicoid, we have remarked that for constant v we have straight lines $u \mapsto (u \cos v, u \sin v, bv)$, so

$$\frac{\partial \Phi}{\partial u} = (\cos v, \sin v, 0).$$

Similarly we remarked that for constant u we have helices. We have already computed the tangent vector of a helix, though it is simple enough to repeat it

$$\frac{\partial \Phi}{\partial v} = (-u \sin v, u \cos v, b).$$

With respect to this basis, we have

$$\begin{aligned} g_{1\,1}(u,v) &= \tfrac{\partial \Phi}{\partial u} \cdot \tfrac{\partial \Phi}{\partial u} = 1, \\ g_{1\,2}(u,v) = g_{2\,1}(u,v) &= \tfrac{\partial \Phi}{\partial u} \cdot \tfrac{\partial \Phi}{\partial v} = 0, \\ g_{2\,2}(u,v) &= \tfrac{\partial \Phi}{\partial v} \cdot \tfrac{\partial \Phi}{\partial v} = u^2 + b^2. \end{aligned}$$

Notice that $g_{1\,2}$ is always zero for this example, so the coordinate basis vectors are perpendicular at every point of the surface.

Let us introduce one more tool inspired by the previous sections before we dive into the geometry of surfaces. We saw how useful the Frenet frame was, and it would be nice to have something similar for surfaces. We already have two vectors that span the tangent space, although they are not necessarily unit length or orthogonal. The cross product $\frac{\partial \Phi}{\partial u} \times \frac{\partial \Phi}{\partial v}$ is perpendicular to both coordinate basis vectors and is non-zero because these vectors are linearly independent. We define the *surface normal* ν to the unit length rescaling of this cross product.

Notice that there are two possible unit length vectors perpendicular to the tangent plane. Per our definition, which one we choose comes down to which coordinate we call u and which we call v. This is effectively the choice of an orientation for the surface, and we will refer to this choice of sign of ν as the orientation. Intuitively this can be thought of as choosing one side of the surface to be the 'plus' side. The choice of sign is determined if we think of ν as a function in the coordinates, but it is potentially ambiguous if we think of it as a function from the unparameterised surface Σ. Fortunately, many definitions using ν do not depend on the sign or do so only in a trivial way. Mirroring the notation for curves on the surface, we will use $\nu : \Sigma \to \mathbb{R}^3$ and $\tilde{\nu} = \nu \circ \Phi : U \to \mathbb{R}^3$ for the two closely related functions.

Finally, the output of ν is a unit length vector in $\mathbb{R}^3$. These are the points of the unit sphere $\mathbb{S}^2$. It is customary therefore to write the target of ν as a sphere and not $\mathbb{R}^3$. Particularly when we think of $\nu : \Sigma \to \mathbb{S}^2$ it is common to call it the *Gauss map*.

Example 1.19. We can compute the surface normal of the helicoid

$$\begin{aligned} \tfrac{\partial \Phi}{\partial u} \times \tfrac{\partial \Phi}{\partial v} &= (b \sin v, -b \cos v, u) \\ \left\| \tfrac{\partial \Phi}{\partial u} \times \tfrac{\partial \Phi}{\partial v} \right\|^2 &= b^2 + u^2 \\ \tilde{\nu} = \frac{\frac{\partial \Phi}{\partial u} \times \frac{\partial \Phi}{\partial v}}{\left\| \frac{\partial \Phi}{\partial u} \times \frac{\partial \Phi}{\partial v} \right\|} &= \frac{1}{\sqrt{b^2 + u^2}} (b \sin v, -b \cos v, u). \end{aligned}$$

Notice that this is basically the same formula as the binormal B of the helix. This is because $\frac{\partial \Phi}{\partial u}$ is the negative of the normal of the helix, but we are also taking the cross product in the other order.

The normal, or rather $\frac{\partial \Phi}{\partial u} \times \frac{\partial \Phi}{\partial v}$, also gives a definition of surface area. The example of the Schwarz lantern shows that for a surface unlike a curve, we cannot make a definition just by taking straight line approximations. Instead we approximate a surface by the parallelogram spanned by the coordinate vector basis. The area of this parallelogram is the length of the cross product. Thus we make the definition that the surface area is

$$\text{Area} = \int_U \left\| \tfrac{\partial \Phi}{\partial u} \times \tfrac{\partial \Phi}{\partial v} \right\| \, du \, dv.$$

1.5 Curvatures

But let us turn to our main focus: curvature. We want to use curves on the surface to say something about the curvature of the surface itself. But this is not as simple as it might first appear. Consider the example of a plane. This is a surface that has no curvature (by any reasonable definition). But we can take circles of any size in the plane, and therefore there are curves with any amount of curvature. The question is how to distinguish curvature that arises because of the choice of curve from curvature that is forced by the surface itself. The answer is to define the normal curvature of a curve on a surface and prove Meusnier's theorem.

Definition 1.20. *Let $\alpha = \Phi \circ \tilde{\alpha}$ be a curve on a regular surface Σ with normal ν. Suppose that α is parameterised by arc-length as a curve in $\mathbb{R}^3$. The* normal curvature *is $\kappa_n := \alpha'' \cdot \nu$.*

Recall that for an arc-length parameterised curve $\alpha'' = \kappa N$, where N is the normal of the curve. So the normal curvature is $\kappa \cos\theta$ for θ the angle between the normal of the curve and the normal of the surface. This shows that the normal curvature is at most the curvature. There is also the possibility of a sign, but this depends on the choice of orientation.

Example 1.21. We mentioned already the case of a plane Σ, of which $\Phi(u, v) = (u, v, 0)$ is an example. The normal to the surface is $\nu = (0, 0, 1)$, which is constant with respect to the point on the surface. Since Σ is a plane, it must be the osculating plane of every curve on it. Hence N is always perpendicular to ν, and the normal curvature of every curve is zero.

Example 1.22. Next we consider a sphere of radius R. Let's argue geometrically so we don't have to do any calculations. Consider the curve which is the equator of the sphere. It is a circle in the plane $z = 0$, so this plane is its osculating plane and the circle is its osculating circle. Its curvature is $\kappa = R^{-1}$. The normal of the curve is unit vector that points towards the center of the sphere. The normal of the surface is the unit vector points either towards or away from the center of the sphere. Therefore the normal curvature of the equator is $\pm R^{-1}$. This argument applies not just to the equator, but to any great circle of the sphere.

Example 1.23. Our main example for this chapter is the helicoid, so of course we must examine its normal curvatures. We have considered two special sets of curves on the helicoid: the helices and the radial lines. The radial lines are lines, and so have zero curvature. Thus their normal curvature is also zero. The helix $\alpha_u(t) = (u\cos t, u\sin t, bt)$ has curvature $u(u^2 + b^2)^{-1}$. But the normal of this curve is $-\frac{\partial \Phi}{\partial u}$, as we remarked upon in Example 1.19 and therefore the dot product with the surface normal is zero. This shows that the helices also have zero normal curvature in the helicoid.

Let us examine several curves at once, all of which pass through the point $\Phi(1, 0) = (1, 0, 0)$. Consider

$$\alpha(t) = \Phi(w^1 t + 1, w^2 t) = \Big((w^1 t + 1)\cos(w^2 t), (w^1 t + 1)\sin(w^2 t), b w^2 t\Big).$$

There is not a nice arc-length parameterisation for this curve. Indeed

$$\begin{aligned}\alpha'(t) &= w^1 \tfrac{\partial \Phi}{\partial u} + w^2 \tfrac{\partial \Phi}{\partial v} \\ &= w^1 \Big(\cos(w^2 t), \sin(w^2 t), 0\Big) + w^2 \Big(-(w^1 t + 1)\sin(w^2 t), (w^1 t + 1)\cos(w^2 t), b\Big)\end{aligned}$$

shows that the length of the tangent vector of this curve is quite a complicated function. However, we are mainly interested in the behaviour at the point with $t = 0$, where

$$\alpha'(0) = \Big(w^1, w^2, w^2 b\Big).$$

We can simplify the calculations a little if we choose the constants w^1, w^2 such that this is a unit length. That means $\|\alpha'(0)\|^2 = (w^1)^2 + (w^2)^2(1 + b^2) = 1$. The osculating plane of the curve is spanned by the first and second derivatives, regardless of the parameterisation, since

$$\frac{d^2\alpha}{dt^2} = \frac{d}{dt}\left(\|\alpha'\|\, T\right) = \frac{dt}{ds}\frac{d}{ds}\left(\|\alpha'\|\, T\right) = \|\alpha'\|\left(\frac{d\|\alpha'\|}{ds}\, T + \|\alpha'\|\kappa N\right). \tag{1.24}$$

In fact, with this formula we almost have the answer, because T is orthogonal to the surface normal ν. Taking the dot product on both sides

$$\alpha'' \cdot \nu = \|\alpha'\|^2 \kappa N \cdot \nu = \|\alpha'\|^2 \kappa_n.$$

Thus it only remains to carry out this calculation, using the surface normal N from Example 1.19:

$$\begin{aligned}
\alpha''(t) &= 2w^1w^2\Big(-\sin(w^2t), \cos(w^2t), 0\Big) - (w^2)^2\Big((w^1t+1)\cos(w^2t), (w^1t+1)\sin(w^2t), 0\Big)\\
\alpha''(0) &= \Big(-(w^2)^2, 2w^1w^2, 0\Big)\\
\tilde{\nu}(0,1) &= \frac{1}{\sqrt{1+b^2}}(0, -b, 1)\\
\kappa_n &= \|\alpha'(0)\|^{-2}\alpha''(0)\cdot\tilde{\nu}(0,1) = \frac{1}{\sqrt{1+b^2}}\big(0 - 2bw^1w^2 + 0\big) = \frac{-2b}{\sqrt{1+b^2}}\, w^1w^2.
\end{aligned}$$

This last example shows that the calculation for the normal curvature is not too bad. But we can simplify the calculation in such a way that we don't even have to calculate α''! Consider any curve α on the surface Σ. Because the surface normal is unit length, if we differentiate $\nu(\alpha(t)) \cdot \nu(\alpha(t)) = 1$ we obtain

$$2\nu \cdot \frac{d}{dt}\Big(\nu(\alpha(t))\Big) = 0.$$

Therefore the derivative of the surface normal lies in the tangent plane. First we do a little parameter function shuffle: $\nu \circ \alpha = \nu \circ \Phi \circ \tilde{\alpha} = \tilde{\nu} \circ \tilde{\alpha}$. Now when we apply the chain rule the 'middle' stage is the coordinate chart $U \subset \mathbb{R}^2$:

$$\frac{d}{dt}\Big(\nu(\alpha(t))\Big) = \frac{d}{dt}\Big(\tilde{\nu}(\tilde{\alpha}(t))\Big) = \frac{\partial\tilde{\nu}}{\partial u}\frac{d\tilde{\alpha}^1}{dt} + \frac{\partial\tilde{\nu}}{\partial v}\frac{d\tilde{\alpha}^2}{dt}$$

Likewise we know that along the curve $\alpha'(t) \cdot \nu(\alpha(t)) = 0$, so we can differentiate this relation

and obtain

$$\begin{aligned}\kappa_n &= \|\alpha'\|^{-2}\,\alpha''\cdot\nu = -\|\alpha'\|^{-2}\,\alpha'\cdot\frac{d}{dt}\Big(\nu(\alpha(t))\Big)\\ &= -\|\alpha'\|^{-2}\left(\frac{\partial\Phi}{\partial u}\frac{d\tilde\alpha^1}{dt}+\frac{\partial\Phi}{\partial v}\frac{d\tilde\alpha^2}{dt}\right)\cdot\left(\frac{\partial\tilde\nu}{\partial u}\frac{d\tilde\alpha^1}{dt}+\frac{\partial\tilde\nu}{\partial v}\frac{d\tilde\alpha^2}{dt}\right)\\ &= \|\alpha'\|^{-2}\begin{pmatrix}\frac{d\tilde\alpha^1}{dt} & \frac{d\tilde\alpha^2}{dt}\end{pmatrix}\begin{pmatrix}-\frac{\partial\Phi}{\partial u}\cdot\frac{\partial\tilde\nu}{\partial u} & -\frac{\partial\Phi}{\partial u}\cdot\frac{\partial\tilde\nu}{\partial v}\\ -\frac{\partial\Phi}{\partial v}\cdot\frac{\partial\tilde\nu}{\partial u} & -\frac{\partial\Phi}{\partial v}\cdot\frac{\partial\tilde\nu}{\partial v}\end{pmatrix}\begin{pmatrix}\frac{d\tilde\alpha^1}{dt}\\ \frac{d\tilde\alpha^2}{dt}\end{pmatrix}\end{aligned}$$

This matrix defines the *second fundamental form* on the tangent plane of the surface, notated with $\mathrm{I\!I}$ or h. We have proved

Theorem 1.25 (Meusnier)**.** *All curves on a regular surface Σ having at some point the same tangent vector w have at that point the same normal curvature. Their normal curvature is given by*

$$\kappa_n = g(w,w)^{-1}h(w,w).$$

Because of this theorem it makes sense speak of the normal curvature $\kappa_n(w)$ of a surface in a direction w. For this reason we say that the normal curvature is telling us something about the curvature of the surface itself, rather than the curves on the surface.

Exercise 1.26. Prove the follow corollaries of Meusnier's theorem. Try as much as possible to argue geometrically rather than relying on calculation. Fix a tangent vector w to the regular surface Σ and let κ_n be the normal curvature in this direction.

a. Consider the plane P spanned by w and ν. Argue that the intersection $\Sigma\cap P$ is a curve with tangent vector w whose curvature is $|\kappa_n|$.

b. Extend this argument to show that for every $\kappa\geq|\kappa_n|$ there is a curve α_κ with tangent vector w and curvature κ.

c. Argue that the union of the osculating circles of every curve with tangent vector w form a sphere with radius κ_n^{-1}.

Let us investigate the second fundamental form a little more. We know that $\frac{\partial\Phi}{\partial u}\cdot\tilde\nu=\frac{\partial\Phi}{\partial v}\cdot\tilde\nu=0$. Differentiating these with respect to the coordinates u and v give the relations

$$\begin{aligned}\frac{\partial^2\Phi}{\partial u\partial u}\cdot\tilde\nu+\frac{\partial\Phi}{\partial u}\cdot\frac{\partial\tilde\nu}{\partial u}&=0 & h_{1\,1}&=\frac{\partial^2\Phi}{\partial u\partial u}\cdot\tilde\nu,\\ \frac{\partial^2\Phi}{\partial u\partial v}\cdot\tilde\nu+\frac{\partial\Phi}{\partial u}\cdot\frac{\partial\tilde\nu}{\partial v}&=0 & h_{1\,2}&=\frac{\partial^2\Phi}{\partial u\partial v}\cdot\tilde\nu,\\ \frac{\partial^2\Phi}{\partial v\partial u}\cdot\tilde\nu+\frac{\partial\Phi}{\partial v}\cdot\frac{\partial\tilde\nu}{\partial u}&=0 & h_{2\,1}&=\frac{\partial^2\Phi}{\partial u\partial v}\cdot\tilde\nu=h_{1\,2},\\ \frac{\partial^2\Phi}{\partial v\partial v}\cdot\tilde\nu+\frac{\partial\Phi}{\partial v}\cdot\frac{\partial\tilde\nu}{\partial v}&=0 & h_{2\,2}&=\frac{\partial^2\Phi}{\partial v\partial v}\cdot\tilde\nu.\end{aligned}$$

Not only do these relations give an easier method to calculate h, but they show that the second fundamental form is a symmetric bilinear form. We have already seen in Example 1.23 that the normal curvature can be both positive and negative for different directions at the same point, therefore the second fundamental form is not positive definite in general.

Example 1.27. We use these formulas to calculate the second fundamental form of the helicoid. Recall

$$\begin{aligned}\tfrac{\partial\Phi}{\partial u} &= \Big(\cos v, \sin v, 0\Big)\\ \tfrac{\partial\Phi}{\partial v} &= \Big(-u\sin v, u\cos v, b\Big)\\ \tilde{\nu} &= \frac{1}{\sqrt{b^2+u^2}}(b\sin v, -b\cos v, u)\end{aligned}$$

so

$$\begin{aligned}\tfrac{\partial^2\Phi}{\partial u\partial u} &= \Big(0,0,0\Big)\\ \tfrac{\partial^2\Phi}{\partial u\partial v} &= \Big(-\sin v, \cos v, 0\Big)\\ \tfrac{\partial^2\Phi}{\partial v\partial v} &= \Big(-u\cos v, -u\sin v, 0\Big).\end{aligned}$$

Hence

$$\begin{aligned}h_{1\,1} &= \frac{\partial^2\Phi}{\partial u\partial u}\cdot\tilde{\nu} = 0\\ h_{1\,2} = h_{2\,1} &= \frac{\partial^2\Phi}{\partial u\partial v}\cdot\tilde{\nu} = \frac{-b}{\sqrt{b^2+u^2}}\\ h_{2\,2} &= \frac{\partial^2\Phi}{\partial v\partial v}\cdot\tilde{\nu} = 0.\end{aligned}$$

The fact that $h_{1\,1} = h_{2\,2} = 0$ explains the behaviour in Example 1.23 that the normal curvature in the direction of the coordinate basis vectors was zero. Indeed, we easily reproduce the result from that example for any point on the helicoid not just $(1,0,0)$:

$$\kappa_n = g(w,w)^{-1}h(w,w) = 1\begin{pmatrix}w^1 & w^2\end{pmatrix}\begin{pmatrix}0 & \frac{-b}{\sqrt{b^2+u^2}}\\ \frac{-b}{\sqrt{b^2+u^2}} & 0\end{pmatrix}\begin{pmatrix}w^1\\ w^2\end{pmatrix} = \frac{-2b}{\sqrt{b^2+u^2}}\,w^1w^2$$

for all w with $g(w,w) = (w^1)^2 + (u^2+b^2)(w^2)^2 = 1$.

The formula for the normal curvature shows that it is invariant under rescaling of the tangent vector w. The geometric explanation is of course that the normal curvature was defined using an arc-length parameterised curve. This motivates us to consider at each point of the surface the normal curvature as a function of unit-length tangent vectors. The set of unit-length tangent vectors is a circle in the tangent plane. So we have a continuous function from a circle to $\mathbb{R}$, hence it must have a maximum and minimum. Actually this function is extremely well-behaved

Theorem 1.28 (Euler)**.** *Fix a point on a regular surface Σ and let κ_1 be the maximum of the normal curvatures and κ_2 the minimum. Let e_1 be a unit-length tangent vector such that $\kappa_n(e_1) = \kappa_1$. Let e_2 be a unit-length tangent vector perpendicular to e_1. Then for any unit-length tangent vector $w = \cos\varphi\, e_1 + \sin\varphi\, e_2$ the normal curvature is*

$$\kappa_n(w) = \kappa_1\cos^2\varphi + \kappa_2\sin^2\varphi.$$

There are some special cases to observe. If the normal curvature is constant at a point, such as

for the sphere, then this formula still works, as $\kappa_1 \cos^2\varphi + \kappa_1 \sin^2\varphi = \kappa_1$. Points where this is the case are called *umbilic points*. For umbilic points, every direction can be chosen for e_1. If the normal curvature is not constant at a point, then we see that the maximum and minimum occur exactly twice, $\varphi = 0, \pi$ for the maximum and $\varphi = \pi/2, 3\pi/2$ for the minimum. Essentially there is one maximum and one minimum direction (up to sign) and they are perpendicular to one another. They are called the *principal directions* and κ_1, κ_2 are call the *principal curvatures*.

Proof. Let e_1, e_2, w be as in the statement of the theorem. We expand using bilinearity and symmetry

$$h(w,w) = \kappa_1 \cos^2\varphi + 2h(e_1,e_2)\cos\varphi\sin\varphi + h(e_2,e_2)\sin^2\varphi.$$

Taking the derivative with respect to φ gives

$$\left.\frac{dh(w,w)}{d\varphi}\right|_{\varphi=0} = 0 + 2h(e_1,e_2)(0+1) + 0.$$

Thus e_1 is a maximum point only if $h(e_1,e_2) = 0$. Further

$$h(w,w) = \kappa_1\cos^2\varphi + h(e_2,e_2)\sin^2\varphi \geq h(e_2,e_2)\cos^2\varphi + h(e_2,e_2)\sin^2\varphi = h(e_2,e_2).$$

again using that κ_1 is the maximum value. This shows that $h(e_2,e_2) = \kappa_2$ is the minimum normal curvature. □

Remark 1.29. The proof shows, particularly the part where $h(e_1,e_2) = 0$, that in this orthonormal basis e_1, e_2 that the second fundamental form is diagonalised with the principal curvatures on the diagonal.

Remark 1.30. Suppose that you have a curve α in a principal direction and you investigate the derivative of the surface normal $\nu(\alpha(t))$. Then it transpires that this derivative, which must lie in the tangent plane, is in the principal direction. This is an alternative method of characterising principal directions, and perhaps the more common one in textbooks.

Example 1.31. For the plane and the sphere of radius R, the normal curvature is constant and equal at every point, respectively 0 and R^{-1}.

Example 1.32. For the helix we have already computed the normal curvature function, but we have not determined the principal curvatures and directions. Write a unit-length vector as $w^1 = \cos\phi, w^2 = \frac{1}{\sqrt{u^2+b^2}}\sin\phi$ with respect to the coordinate vector basis. Then

$$\kappa_n = \frac{-2b}{u^2+b^2}\cos\phi\sin\phi = \frac{-b}{u^2+b^2}\sin(2\phi) = \frac{b}{u^2+b^2}\cos(2\varphi) = \frac{b}{u^2+b^2}\Big(\cos^2\varphi - \sin^2\varphi\Big),$$

for $\phi = \varphi - \pi/4$. Therefore we see that the principal curvatures of the helicoid are $\pm\frac{b}{u^2+b^2}$, and the principal directions are half-way between the helix and radial directions.

If this were a course purely about curves and surfaces, we would spend a lot more time here investigating special surfaces. For example, the only surfaces where every point is umbilic are (all or parts of) the plane or the sphere. One can also try to find curves on the surface whose tangent vector is always a principal direction, a so-called *line of curvature*. And so on. This

would lead us naturally to the definitions of elliptic and hyperbolic points of a surface, and to Gauss curvature. Instead we will give definitions and claim some properties before moving on.

From Euler's theorem, we see that the normal curvatures at a point are completely characterised by the principal curvatures. From these we define two types of curvature. Conversely, if we know the Gauss and mean curvatures, it is possible to solve for the principal curvatures. Thus the normal curvatures at a point are equivalently described by these two quantities.

Definition 1.33. *The* Gauss curvature *of a surface at a point is* $K = \kappa_1\kappa_2$ *and the* mean curvature *is* $H = \frac{1}{2}(\kappa_1 + \kappa_2)$.

We include a useful formula.

Lemma 1.34. *Let* v, w *are an orthonormal basis for* T_pM *and write* h *with respect to this basis. Then the Gauss curvature at* p *is* $h_{1\,1}h_{2\,2} - h_{1\,2}^2$.

Proof. Be careful: this looks like a determinant but the formula is only true for an orthonormal basis. Although the determinant of the matrix of a linear map is basis-independent, this is not true of the matrix of a bilinear form.

Any unit-length vector of T_pM can be written as $v\cos\theta + w\sin\theta$. Meusnier's theorem and bilinearity tells us

$$\begin{aligned}\kappa_n &= h_{11}\cos^2\theta + 2h_{12}\cos\theta\sin\theta + h_{22}\sin^2\theta \\ &= h_{11}\frac{1+\cos 2\theta}{2} + h_{12}\sin 2\theta + h_{22}\frac{1-\cos 2\theta}{2} \\ &= \frac{1}{2}(h_{11}+h_{22}) + \frac{1}{2}(h_{22}+h_{11})\cos 2\theta + h_{12}\sin 2\theta \\ &= \frac{1}{2}(h_{11}+h_{22}) + R\cos 2\theta,\end{aligned}$$

for $R^2 = \frac{1}{4}(h_{22}+h_{11})^2 + h_{12}^2$. We see that this obtains its maximum and minimum for $\cos 2\theta = \pm 1$. Hence

$$K = \kappa_1\kappa_2 = \frac{1}{4}(h_{11}+h_{22})^2 - R^2 = h_{11}h_{22} - h_{12}^2. \qquad \square$$

Despite the fact that these our formulas for these quantities depend on the first and second fundamental form, it turns out that it is possible to calculate the Gauss curvature using only the first fundamental form. We say that the Gauss curvature is intrinsic to the surface, whereas the mean curvature is extrinsic. This is known as Gauss' *Theorem Egregium* (the extraordinary theorem) and we will prove a generalisation of it in Chapter 5. In fact, we are finally in a position where we can give good definitions of intrinsic and extrinsic.

Definition 1.35. *Suppose we have two parameterised regular surfaces* Φ *and* $\tilde{\Phi}$ *from the same open subset* $U \subset \mathbb{R}^2$ *to* $\mathbb{R}^3$. *If both first fundamental forms* $g_{i\,j}, \tilde{g}_{i\,j}$, *considered as functions on* U, *are equal, then we say that the surfaces are* isometric. *Geometrically, this means that the distances between corresponding points on both surfaces are equal. Quantities that depend only on the first fundamental form are said to be intrinsic.*

Example 1.36. The classic example of an isometric transformation is rolling a sheet of paper into a cylinder. Consider $U = (-\pi, \pi) \times \mathbb{R}$ and the two parameterisations $\Phi(u, v) = (1, u, v)$ and $\tilde{\Phi}(u, v) = (\cos u, \sin u, v)$. The domain has been chosen so that both parameterisations are injective, as we require. We compute the first fundamental forms

$$\begin{aligned} \tfrac{\partial \Phi}{\partial u} = (0, 1, 0),\ \tfrac{\partial \Phi}{\partial v} = (0, 0, 1) \qquad & g_{1\,1} = 1,\ g_{1\,2} = g_{2\,1} = 0,\ g_{2\,2} = 1, \\ \tfrac{\partial \tilde{\Phi}}{\partial u} = (-\sin u, \cos u, 0),\ \tfrac{\partial \tilde{\Phi}}{\partial v} = (0, 0, 1) \qquad & \tilde{g}_{1\,1} = 1,\ \tilde{g}_{1\,2} = \tilde{g}_{2\,1} = 0,\ \tilde{g}_{2\,2} = 1. \end{aligned}$$

So these are indeed isometric surfaces.

From Example 1.21 we know that the normal curvature of the plane is identically zero. Thus so too are the Gauss and mean curvatures.

For the cylinder, we compute the normal and second fundamental form

$$\begin{gathered} \nu = (\cos u, \sin u, 0),\ \tfrac{\partial^2 \tilde{\Phi}}{\partial u^2} = (-\cos u, -\sin u, 0),\ \tfrac{\partial^2 \tilde{\Phi}}{\partial v \partial u} = \tfrac{\partial^2 \tilde{\Phi}}{\partial v^2} = 0 \\ \tilde{h}_{1\,1} = -1,\ \tilde{h}_{1\,2} = \tilde{h}_{2\,2} = 0. \end{gathered}$$

This is already diagonalised, so we see that the principal curvatures are -1 and 0. (The minus sign is because our cylinder has the outward pointing normal ν but the curve in the surface have an inward pointing normal N.) The Gauss curvature is everywhere zero, same as the plane, but the mean curvature is everywhere $-\frac{1}{2}$.

1.6 Minimal Surfaces

In the final section of this chapter we will indulge my tastes and look at a class of special surfaces related to my own research. A minimal surface is one that has the smallest surface area for a given boundary. To simplify matters, we will consider only graphs. Let U be a bounded region of the plane with smooth boundary, and let $g : \partial U \to \mathbb{R}$ be a continuous function. We consider the set of functions

$$\mathcal{F}_g = \{f \in \mathcal{C}^2(\overline{U}) \mid f|_{\partial U} = g\},$$

and their graphs $\Phi_f(u, v) = (u, v, f(u, v))$. Graphs are always regular surfaces. The area is

$$\text{Area}(f) = \int_U \| \tfrac{\partial \Phi}{\partial u} \times \tfrac{\partial \Phi}{\partial v} \| \, du\, dv = \int_U \| \left(\tfrac{\partial f}{\partial u}, \tfrac{\partial f}{\partial v}, 1 \right) \| \, du\, dv = \int_U \sqrt{\left(\tfrac{\partial f}{\partial u}\right)^2 + \left(\tfrac{\partial f}{\partial v}\right)^2 + 1} \, du\, dv.$$

If a surface is a minimal surface, then it must be a critical point for the area. That means, for all variations that don't change the boundary $h \in \mathcal{F}_0$ we have

$$\frac{d}{ds}\text{Area}(f + sh)\bigg|_{s=0} = 0.$$

We compute

$$\begin{aligned}
&\frac{d}{ds}\text{Area}(f + sh)\bigg|_{s=0} = \int_U \frac{d}{ds}\sqrt{(\partial_u f + s\partial_u h)^2 + (\partial_v f + s\partial_v h)^2 + 1}\bigg|_{s=0} du\, dv \\
&= \int_U \frac{\partial_u f \partial_u h + \partial_v f \partial_v h}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}} \, du\, dv \\
&= \int_U \frac{\partial_u f}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}} \, \partial_u h \, du\, dv + \int_U \frac{\partial_v f}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}} \, \partial_v h \, du\, dv \\
&= -\int_U \frac{\partial}{\partial u}\left(\frac{\partial_u f}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}} \right) h \, du\, dv - \int_U \frac{\partial}{\partial v}\left(\frac{\partial_v f}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}} \right) h \, du\, dv,
\end{aligned}$$

using partial integration and the fact that h is zero on the boundary. The only way that this can be zero for all $h \in \mathcal{F}_0$ is if

$$\frac{\partial}{\partial u}\left(\frac{\partial_u f}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}} \right) + \frac{\partial}{\partial v}\left(\frac{\partial_v f}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}} \right) = 0.$$

This can also be written in vector notation as

$$\nabla \cdot \left(\frac{\nabla f}{\|\nabla f\|^2 + 1} \right) = 0. \tag{1.37}$$

This is called the minimal graph equation, which is treated in the course *Partial Differential Equations*. If you expand out the derivatives you obtain

$$\begin{aligned}
&\frac{\partial_u^2 f + \partial_v^2 f}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}} - \frac{1}{2}\frac{\partial_u f \, (2\partial_u f \partial_u^2 f + 2\partial_v f \partial_u \partial_v f) + \partial_v f \, (2\partial_u f \partial_u \partial_v f + 2\partial_v f \partial_v^2 f)}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}^3} \\
&= \frac{(\partial_u^2 f + \partial_v^2 f)((\partial_u f)^2 + (\partial_v f)^2 + 1) - \partial_u f \, (\partial_u f \partial_u^2 f + \partial_v f \partial_u \partial_v f) - \partial_v f \, (\partial_u f \partial_u \partial_v f + \partial_v f \partial_v^2 f)}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}^3} \\
&= \frac{\partial_u^2 f((\partial_v f)^2 + 1) + \partial_v^2 f((\partial_u f)^2 + 1) - 2\partial_u f \partial_v f \partial_u \partial_v f}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}^3}.
\end{aligned}$$

We pause our calculation here to compute the mean curvature of a graph. As in the previous section, we need the second derivatives of the parameterisation

$$\frac{\partial^2 f}{\partial u^2} = (0,0,\partial_u^2 f),\ \frac{\partial^2 f}{\partial u \partial v} = (0,0,\partial_u\partial_v f),\ \frac{\partial^2 f}{\partial v^2} = (0,0,\partial_v^2 f),$$
$$h = \frac{1}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}} \begin{pmatrix} \partial_u^2 f & \partial_u\partial_v f \\ \partial_u\partial_v f & \partial_v^2 f \end{pmatrix}$$

Unfortunately, there is no reason that this will be a diagonal matrix in general, so how should we find the principal curvatures? We will use a little bit of linear algebra. We define a linear transformation A on the tangent space using the formula $h(v,w) = g(v, Aw)$. This is well-defined because g is positive definite. If we use the basis e_1, e_2 of principal directions then we have

$$0 = h(e_1, e_2) = g(e_1, Ae_2), \qquad 0 = h(e_2, e_1) = g(e_2, Ae_1).$$

We conclude that Ae_2 is orthogonal to e_1. Therefore it must be a multiple λ_2 of e_2. Likewise $Ae_1 = \lambda_1 e_1$. Further

$$\kappa_i = h(e_i, e_i) = g(e_i, Ae_i) = \lambda_i g(e_i, e_i) = \lambda_i.$$

In other words, the principal curvatures are the eigenvalues of this matrix A. Hence the mean curvature is $H = \frac{1}{2}(\kappa_1 + \kappa_2) = \frac{1}{2}\operatorname{tr} A$. In terms of the coordinate basis, the matrix A is the product of the inverse of g with h.

$$\begin{aligned} A &= \begin{pmatrix} 1+(\partial_u f)^2 & \partial_u f \partial_v f \\ \partial_u f \partial_v f & 1+(\partial_v f)^2 \end{pmatrix}^{-1} \frac{1}{\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}} \begin{pmatrix} \partial_u^2 f & \partial_u\partial_v f \\ \partial_u\partial_v f & \partial_v^2 f \end{pmatrix} \\ &= \frac{1}{\det(g)\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}} \begin{pmatrix} 1+(\partial_v f)^2 & -\partial_u f \partial_v f \\ -\partial_u f \partial_v f & 1+(\partial_u f)^2 \end{pmatrix} \begin{pmatrix} \partial_u^2 f & \partial_u\partial_v f \\ \partial_u\partial_v f & \partial_v^2 f \end{pmatrix} \\ &= \frac{1}{*} \begin{pmatrix} (1+(\partial_v f)^2)\partial_u^2 f - \partial_u f \partial_v f \partial_u\partial_v f & * \\ * & -\partial_u f \partial_v f \partial_u\partial_v f + (1+(\partial_u f)^2)\partial_v^2 f \end{pmatrix}, \end{aligned}$$

where we have used $*$ to abbreviate expressions that are too long and not important. Hence

$$H = \frac{1}{2}\operatorname{tr} A = \frac{(1+(\partial_v f)^2)\partial_u^2 f - 2\partial_u f \partial_v f \partial_u\partial_v f + (1+(\partial_u f)^2)\partial_v^2 f}{\det(g)\sqrt{(\partial_u f)^2 + (\partial_v f)^2 + 1}}.$$

But notice, the numerator of H is exactly the same as the numerator in the minimal graph equation. Thus one is zero if and only if the other is too. In summary, a surface solves the minimal graph equation if and only if it has zero mean curvature.

We can turn this around to give a geometric characterisation of mean curvature: it is a measure of how far a surface deviates from being locally area minimising. For example, the plane has zero mean curvature because if you draw a loop on the plane, the least area surface with that boundary is a plane. The same is true if you draw a small circle on a sphere: a flat circle would have less area then the spherical cap. This is a geometric argument that a sphere has non-zero mean curvature.

Remarkably, the helicoid is a minimal surface! We have seen in Example 1.32 that the principal curvatures are $\pm\frac{b}{u^2+b^2}$. Therefore the mean curvature is zero. Minimal surfaces have a rich and fascinating theory. Just one example would be that the helicoid along with the catenoid belongs to a family of minimal surfaces, all of which are (locally) isometric to one another.

Chapter 2

Manifolds

In this chapter we give the 'patchwork' definition of manifolds. Manifolds are geometrically nice spaces and a natural generalisation of $\mathbb{R}^n$. The most common way to define a manifold is as a special type of topological space, namely a 'second-countable Hausdorff locally-euclidean topological space with an atlas'[1]. Because students usually encounter differential geometry before abstract topology, the lecturer then gives a speed run of all the definitions in topology. I think this approach is better suited to the second time you encounter manifolds. Then you already know a little bit about what makes manifolds nice, and you can appreciate the interesting but weird topological spaces that need to be excluded from the definition. The standard approach in effect defines a manifold by saying what a manifold isn't. In the approach below we avoid defining general topological spaces and instead use concrete gluing construction of open sets of $\mathbb{R}^n$. After this construction we will still need to impose certain conditions, so topology cannot be avoided completely, but hopefully they are suitable for a new-commer to manifolds.

Before we dive into theory, we define a concept that you probably know but have never had a word for. A *partial function* from X to Y is a function from a subset $S \subset X$ to Y. In the context of partial functions, a function with $S = X$ is called a *total function*. Many common functions are really partial functions: $\frac{1}{x}$ and $\sqrt{x}$ are partial functions from $\mathbb{R}$ to $\mathbb{R}$, with S being respectively $\mathbb{R} \setminus \{0\}$ and $[0, \infty)$. There doesn't seem to be good standard terminology to talk about X and S, though S is often called the natural domain. Let's call X the *source* of the function and S the *domain*, with the symbols $\operatorname{src} f = X$ and $\operatorname{dom} f = S$. You are no doubt familiar with the difference between the *codomain* $\operatorname{codom} f = Y$ (also called the target) and the *image* $\operatorname{img} f = f[S]$ (also called the range). We will use $f : X \rightharpoonup Y$ for partial functions (harpoon arrow), in contrast to $f : X \to Y$ for total functions.

Many students in Analysis I are confused about the relationship between injective and surjective, and those students are correct to be confused. Just as surjective means that the image is equal to the codomain, total means that the domain is equal to the source; they are the true counterparts to one another. In fact a partial function that is injective has an inverse partial function. If $f : X \rightharpoonup Y$ is injective then

$$f^{-1}(y) = x \text{ if } y = f(x)$$

is a perfectly valid definition of a partial function $f^{-1} : Y \rightharpoonup X$ with $\operatorname{dom} f^{-1} = \operatorname{img} f$ and

[1] Sources differ: some use second-countable, others use the 'Lindelöff' property, which is equivalent in this context, and others use the more general 'paracompact'

$\operatorname{img} f^{-1} = \operatorname{dom} f$.

Besides inverses, many of the usual definitions for functions carry over with sensible modifications. For example, partial functions from X to Y are equal if they have the same domains and are equal on all inputs. Likewise the composition of two partial functions $f : X \rightharpoonup Y$ and $g : Y \rightharpoonup Z$ is a partial function $g \circ f : X \rightharpoonup Z$, but the domain of $g \circ f$ will be smaller than of the domain of f if the image of f lies partly outside the domain of g. We should also think about the 'empty' partial function. According to the set theory definition of a function, there is exactly one function from the empty set to any other set. This is usually considered a quirk in the definition. But the composition of partial functions where $\operatorname{img} f \cap \operatorname{dom} g = \emptyset$ results in the empty function, so we cannot ignore it.

Remark 2.1. In this course, functions and partial functions will always have open sets as their sources and their domains will be open subsets, unless specifically noted otherwise. This is needed so that derivatives can be defined.

2.1 Manifolds

Fix a dimension n. To keep track of the pieces we will glue together, let us introduce an index set $\mathcal{I}$. In our examples this will usually be a finite set, but we do not make this assumption generally. For every $i \in \mathcal{I}$ let U_i be an open subset of $\mathbb{R}^n$. U_i is called a *chart*. These are the pieces we will glue together.

For two charts, we describe how to glue them together using a partial function. A gluing function $\varphi_{ij} : U_i \rightharpoonup U_j$ is a partial function that is a homeomorphism from its domain to its image. This means it is a homeomorphism (bijective, continuous, continuous inverse) between open subsets $V_i \subset U_i$ and $V_j \subset U_j$. The idea is that the point $x \in V_i$ is glued to $\varphi_{ij}(x) \in V_j$. Note the order of the subscripts: "from i to j". Other names for gluing functions include 'transition functions', 'change of coordinate functions', and 'overlap functions'. We allow here the possibility that $V_i = V_j = \emptyset$ and $\varphi_{ij} : \emptyset \to \emptyset$ is the empty function, this represents the situation that U_i and U_j are not glued together at all. To avoid having too many named sets, we will mostly use $\operatorname{dom} \varphi_{ij}$ instead of V_i.

This information tells how to glue the pieces together, but how should we represent the completed glued object? First we define the *disjoint union*

$$\coprod_{i \in \mathcal{I}} U_i = \{(i, x) \mid i \in \mathcal{I}, x \in U_i\},$$

which is a set of pairs. We think of this as saying that even if a point is common to both U_i and U_j, in the disjoint union we consider it as two separate points. For example, if $U_1 = (-1, 1)$ and $U_2 = (0, 2)$ then the normal union is $U_1 \cup U_2 = (-1, 2)$ but the disjoint union is two intervals. We often do not write the index i if it is clear, and even when it is not clear we tend to write it as a subscript. Continuing the example, $U_1 \coprod U_2$ has two points that might both be called 0.5, namely $0.5_1 \in U_1$ and $0.5_2 \in U_2$. Formally these points should be written $(1, 0.5)$ and $(2, 0.5)$ respectively.

We want to create an equivalence relation on the disjoint union of all the charts such that $x \in U_i \sim y \in U_j$ iff $\varphi_{ij}(x) = y$. If this is to be an equivalence relation, the set of gluing

functions is required to have certain properties. To get reflexivity of $\sim$, we need $\varphi_{i\,i} = \mathrm{id}_{U_i}$. Symmetry of the relation holds if and only if $\varphi_{j\,i} = \varphi_{i\,j}^{-1}$. These are simple enough, but expressing the condition for transitivity is more difficult. The usual way to express the transitivity condition is that $y \sim x, x \sim z \Rightarrow y \sim z$, but if we have symmetry then this is equivalent to $x \sim y, x \sim z \Rightarrow y \sim z$. That means for all $x \in \operatorname{dom}\varphi_{i\,j} \cap \operatorname{dom}\varphi_{i\,k}$ we need $y = \varphi_{i\,j}(x) \in \operatorname{dom}\varphi_{j\,k}$ and $z = \varphi_{i\,k}(x) = \varphi_{j\,k}(y)$. This is usually shorthanded to

$$\varphi_{i\,k} = \varphi_{j\,k} \circ \varphi_{i\,j},$$

but there needs to be some restrictions to make this rigorous. If a set of transitions have these three properties, and thus defines an equivalence relation, we say that fulfill the *cocycle conditions.*

We call $M = \coprod U_i / \sim$ the glued space.[2] Each 'point' in M is an equivalence class of points in different U_i. It is useful to have functions that allow us to move between U_i and M. There is of course the canonical projection $\pi_M : \coprod U_i \to M$ that sends every element to its equivalence class, but this is too rough. We call the restriction $\Phi_i := \pi_M|_{U_i} : U_i \to M$ a parameterisation of $\pi_M[U_i] \subset M$. It sends a point of U_i to its equivalence class in M. We really should think of this as a parameterisation because an ordinary set U_i in euclidean space is describing part of a complicated object M. In the other direction $\phi_i := (\Phi_i)^{-1} : \pi_M[U_i] \to U_i \subset \mathbb{R}^n$ is called a coordinate function. It sends an equivalence class to its representative that lies in U_i. We usually write ϕ_i^{-1} for the parameterisation rather than Φ_i, as it is unnecessary to have two symbols.

Definition 2.2. *An* atlas *is a tuple $\mathcal{A} = (n, \mathcal{I}, \{U_i\}_{i\in\mathcal{I}}, \{\varphi_{i\,j}\}_{i,j\in\mathcal{I}})$, where $\{\varphi_{i\,j}\}_{i,j\in\mathcal{I}}$ fulfills the cocycle conditions. We have included n in our definition of atlas to make all the charts have the same dimension (some authors allow spaces with a mix of dimensions).*

Example 2.3. Take any open subset $U \subset \mathbb{R}^n$. We can construct the trivial atlas for U as follows. Let the index set $\mathcal{I} = \{0\}$ and $U_0 = U$ the only chart. Then $\varphi_{0\,0} = \mathrm{id}_U$ is a gluing function. The cocycle condition is fulfilled, so $(n, \mathcal{I}, \{U\}, \{\varphi_{0\,0}\})$ is an atlas. The corresponding equivalence relation $\sim$ is the weakest equivalence relation on U, namely x is equivalent to itself but no other points. Therefore we say $M = U$.

Example 2.4. Consider a continuous function $g : U \subset \mathbb{R}^n \to \mathbb{R}^m$ and the graph

$$\operatorname{graph} g = \{(x, g(x)) \in \mathbb{R}^{n+m} \mid x \in U\}.$$

Like the previous example, we take $\mathcal{I} = \{0\}$ and $U_0 = U$. The only gluing function is trivial, so on an abstract level there is no difference between U and M. But we want to think of $M = \operatorname{graph} g$. We can do this using coordinates and parametrisations. Here we have the coordinate function $\phi_0 : M \to U_0$ given by $\phi_0(x, g(x)) = x$ and the parameterisation $\phi_0^{-1} : U_0 \to M$ given by $\phi_0^{-1}(x) = (x, g(x))$.

Example 2.5. Consider $n = 1$, $\mathcal{I} = \{1, 2\}$ and $U_1 = (-1, 1), U_2 = (-1, 1)$. To give an atlas it is sufficient to describe $\varphi_{1\,2}$, because the cocycle condition requires $\varphi_{1\,1} = \mathrm{id}_{U_1}$ and $\varphi_{2\,2} = \mathrm{id}_{U_2}$, as well as $\varphi_{2\,1} = \varphi_{1\,2}^{-1}$. Set $V_1 = U_1 \setminus \{0\}$ and $V_2 = U_2 \setminus \{0\}$ and give the gluing

[2] If you are familiar with topology, the quotient topology makes M a locally euclidean topological space.

function $\varphi_{1\,2} : V_1 \to V_2$ the formula

$$\varphi_{1\,2}(x) = \begin{cases} x+1 & \text{for } x \in (-1,0), \\ x-1 & \text{for } x \in (0,1). \end{cases}$$

For example, this tells us that we should glue $-0.5_1 \in U_1$ to $\varphi_{1\,2}(-0.5) = 0.5_2 \in U_2$ and $0.3_1 \in U_1$ to $\varphi_{1\,2}(-0.3) = -0.7_2 \in U_2$. Here the glued space M is a circle, which you can see by cutting two strips of paper of the same length, drawing a number line from -1 to 1 on each of them, and then gluing as instructed. Every point of M is equivalent to a point of U_1 or to 0_2.

Example 2.6. Consider everything the same as in Example 2.5, but this time give $\varphi_{1\,2}$ the formula

$$\varphi_{1\,2}(x) = x, \qquad x \in (-1,0) \cup (0,1).$$

The space M is a called an interval with two origins. This is because every point of M is either $\phi_1^{-1}(x) = \phi_2^{-1}(x)$ for $x \neq 0$, or 0_1, or 0_2.

The interval with two origins might seem like a harmless curiosity, but in fact it is a weird topological space that we want to avoid. Let us say that a sequence p_n in M converges to a point p if the sequence $\phi_i(p_n)$ converges to $\phi_i(p)$ in the chart U_i. In Example 2.6 consider the sequence n^{-1} for $n \in \mathbb{N}^+$. We can view this sequence in U_1 or U_2. In U_1 it has the limit 0_1, but in U_2 is has the limit 0_2. Therefore in M this sequence has two different limits!

Because we want to do calculus, we need to use limits. We are only interested in spaces where sequences have at most one limit. It turns out that the above example is the only way a gluing can cause non-unique limits.

Lemma 2.7 (Non-unique Limits)**.** *A glued space M has non-unique limits if and only if there is a sequence $x_n \in \operatorname{dom}\varphi_{1\,2} \subset U_1$ such that x_n converges to a point $x \in U_1 \setminus \operatorname{dom}\varphi_{1\,2}$ and $\varphi_{1\,2}(x_n)$ converges to a point $y \in U_2$, for some charts U_1, U_2.*

Proof. Suppose first we have a sequence $x_n \in \operatorname{dom}\varphi_{1\,2} \subset U_1$ such that x_n converges to a point $x \in U_1 \setminus \operatorname{dom}\varphi_{1\,2}$ and $\varphi_{1\,2}(x_n)$ converges to a point $y \in U_2$. By definition, x and y are both limits of the sequence x_n. But x is outside the domain of the gluing function, so by definition is it not glued to any point of U_2. Therefore x and y are distinct points in M.

Conversely suppose a glued space M has a sequence with two distinct limits. In euclidean space limits are unique, so if there are two limits then they must come from two charts U_1 and U_2. Let's use the notation $x_n, x \in U_1$ and $y_k, y \in U_2$ with $y_n = \varphi_{1\,2}(x_n)$. Suppose x were in the domain of the gluing function. The gluing function by definition is a continuous function, so we would have $\varphi_{1\,2}(x) = \lim \varphi_{1\,2}(x_n) = \lim y_n = y$. But $\varphi_{1\,2}(x) = y$ means exactly that $x \sim y$ and this contradicts our assumption that the limits are distinct in M. Therefore x is not in the domain of $\varphi_{1\,2}$, but it is the limit of a sequence in the domain. We have shown that $x \in U_1 \setminus \operatorname{dom}\varphi_{1\,2}$. □

Definition 2.8. *We say that an atlas has the* unique limit property *if it does not not satisfy the condition of Lemma 2.7. That is, if for every $i, j \in \mathcal{I}$ and every sequence $x_n \in \operatorname{dom}\varphi_{i\,j}$ that converges to some $x \in U_1 \setminus \operatorname{dom}\varphi_{i\,j}$, the sequence $\varphi_{i\,j}(x_n)$ does not converge in U_2.*

Observe why this is not an issue with Example 2.5. Consider the sequence $x_n = n^{-1}$ in U_1. It has the limit 0_1. On the other hand, the sequence $y_n = \varphi_{1\,2}(x_n) = n^{-1} - 1$ converges to -1 in $\mathbb{R}$, but -1 is not in U_2. Therefore this 'other' limit point is not in the space M.

The other way that gluing can produce a topologically bad space is if we glue too many charts together. We will not provide an example of this; interested students may search for the 'long line' or the 'long ray', which are standard examples of this phenomenon.

Definition 2.9. *A* manifold *is an atlas with the unique limit property and such that the index set is countable.*

It is very common to talk about the glued space M as the manifold without explicitly stating the atlas. This is similar to talking about a vector space as the set of vectors, when in fact it is the operations of addition and scalar multiplication that make a vector space interesting.

There are different sorts of manifolds, based on additional conditions on the gluing functions. We will use the notation that a function is C^ℓ when it is ℓ-times continuously differentiable. By convention, C^0 means that the function is continuous, and C^∞ means that the function is smooth. An atlas (or a manifold) is called C^ℓ when all of the gluing functions are C^ℓ. Probably the most common type of manifold that is studied, and the one we will study in this course, are *smooth* manifolds. Henceforth, when we say manifold we mean smooth manifold.

Because we have given a non-standard definition of manifolds, we should explain how this compares to the standard definition. We do this using the example of a sphere and stereographic projection, which seems to be the first non-trivial example in every book on manifolds.

Example 2.10. The sphere is the set $\mathbb{S}^n = \{p \in \mathbb{R}^{n+1} \mid \|p\| = 1\}$, and we name the north pole $N = (0, \ldots, 0, 1)$ and the south pole $S = (0, \ldots, 0, -1)$. Stereographic projection from the north pole is the function

$$\phi_N : \mathbb{S}^n \setminus \{N\} \to \mathbb{R}^n, \qquad p \mapsto \frac{1}{1 - p^{n+1}}(p^1, \ldots, p^n),$$

and stereographic projection from the south pole is

$$\phi_S : \mathbb{S}^n \setminus \{S\} \to \mathbb{R}^n, \qquad p \mapsto \frac{1}{1 + p^{n+1}}(p^1, \ldots, p^n).$$

These formulas come from the following construction. For any point of the sphere, draw a line in $\mathbb{R}^{n+1}$ to the pole. The result of the stereographic projection is where the line intersects the plane $(y^1, \ldots, y^n, 0)$. Notice that this construction is ill-defined when applied to the pole itself, because a single point does not determine a line, so it is naturally a partial function on the sphere. On the stated domains these are both bijections (in fact they are homeomorphisms). We are more familiar with are the inverse of these functions

$$\phi_N^{-1} : \mathbb{R}^n \to \mathbb{S}^n \setminus \{N\}, \qquad x \mapsto \frac{1}{\|x\|^2 + 1}(2x^1, \ldots, 2x^n, \|x\|^2 - 1),$$

and

$$\phi_S^{-1} : \mathbb{R}^n \to \mathbb{S}^n \setminus \{S\}, \qquad y \mapsto \frac{1}{\|y\|^2 + 1}(2y^1, \ldots, 2y^n, 1 - \|y\|^2).$$

These are regular parametrisations of (parts of) the sphere in the sense of Section 1.4. So given a point p on the sphere, we can apply ϕ_N to get a point in $\mathbb{R}^n$, called its coordinates with respect to ϕ_N, and putting the coordinates into the parameterisation ϕ_N^{-1} gives back the point p. If we have two coordinate functions ϕ_N and ϕ_S and we know the coordinates x with respect to ϕ_N then $\phi_S \circ \phi_N^{-1}(x)$ is the coordinates with respect to ϕ_S. For this reason $\phi_S \circ \phi_N^{-1} : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}^n \setminus \{0\}$ is called the change of coordinates function. In this example, one can calculate that $\varphi_{N\,S}(x) := \phi_S \circ \phi_N^{-1}(x) = \|x\|^{-2}x$.

Now we now see the connection with our definitions above. The index set is $\mathcal{I} = \{N, S\}$. The image of the coordinate functions are the charts, $U_N = \mathbb{R}^n$ and $U_S = \mathbb{R}^n$. And the (non-trivial) gluing function from $U_N \setminus \{0\}$ to $U_S \setminus \{0\}$ is $\varphi_{N\,S}(y) = \|y\|^{-2}y$. We indeed see that the equivalence classes of points are $S = \{0_N\}, N = \{0_S\}$, or $\{y_N, y_S\}$ for $y_S = \|y_N\|^{-2}y_N$ and $y_N \in \mathbb{R}^n \setminus \{0\}$.

Notice that it is not possible to cover the sphere by a single regular parameterisation (parameterisations must be injective). Thus to deal with surfaces generally, on must consider multiple overlapping parameterisations and change of coordinate functions. We avoided this complication in the previous chapter by only working with a piece of the sphere, or working with the sphere geometrically.

Exercise 2.11. Carry out the claimed calculation from the previous example:

$$\varphi_{N\,S}(x) = \phi_S \circ \phi_N^{-1}(x) = \|x\|^{-2}x.$$

In our definition of manifold we began with the charts and gluing function, then constructed M and the coordinate functions. In the standard definition of a manifold you begin with M and coordinate functions, then define the gluing functions as $\varphi_{i\,j} = \phi_j \circ \phi_i^{-1}$. These are equivalent approaches. One of the drawbacks of differential geometry being an old and widely practised field is that notation and terminology has been around for a long time and is not completely standardised. Different authors use the word chart to describe ϕ_i, ϕ_i^{-1}, U_i, or $\phi_i^{-1}[U_i]$, though perhaps calling ϕ_i a chart is the most common usage.

Let us summarise our terminology and the relations between the objects. Charts are open sets of $\mathbb{R}^n$. An atlas is set of charts and gluing functions with the cocycle property. The information of an atlas allows us to glue the charts together to get a manifold M. Functions from the charts to the manifold are called parametrisations and functions from the manifold to the charts are coordinates. A gluing function can also be called a change of coordinate function or a transition function. Conversely, if we start with a space M and coordinates, then the image of coordinate function is a chart and its inverse is a parameterisation. The composition of a parameterisation and a coordinate is a transition function.

2.2 Functions

Next we want to define functions between manifolds. On one hand, there is nothing to do. If we have manifolds M and N, they are sets, and a function $f : M \to N$ is defined in the normal way. But manifolds are more than sets of points, they have atlases. Let

$$\mathcal{A}_M = (n_M, \mathcal{I}_M, \{U_{M\,i}\}_{i\in\mathcal{I}_M}, \{\varphi_{M,i,j}\}_{i,j\in\mathcal{I}_M}),$$
$$\mathcal{A}_N = (n_N, \mathcal{I}_N, \{U_{N\,i}\}_{i\in\mathcal{I}_N}, \{\varphi_{N,i,j}\}_{i,j\in\mathcal{I}_N})$$

be atlases for M and N respectively (the extra subscripts indicate the manifold). Then we can look at f 'in charts'. This means we look at the partial functions

$$f_{i\,k} := \phi_{N\,k} \circ f \circ \phi^{-1}_{M\,i} : U_{M\,i} \rightharpoonup U_{N\,k} \text{ for } i \in \mathcal{I}_M \text{ and } k \in \mathcal{I}_N.$$

These are functions between euclidean spaces, so we can ask whether they are C^ℓ.

Definition 2.12. *A function $f : M \to N$ is C^ℓ at a point $p \in M$ if and there is a chart $U_{M\,i} \ni \phi_{M\,i}(p)$ and chart $U_{N\,k} \ni \phi_{N\,j}(f(p))$ such that $f_{i\,k}$ is C^ℓ, and it obeys a further technical condition.*[3] *We say that f is C^ℓ on M if it is C^ℓ at every point of M.*

The definition of C^ℓ uses a chart. But we know that a point may belong to more than one chart. This opens the possibility that f is C^ℓ at p according to one chart, but not C^ℓ according to another chart. However, because of the relation

$$\begin{aligned} f_{j\,l} &= \phi_{N\,l} \circ f \circ \phi^{-1}_{M\,j} = (\phi_{N\,l} \circ \phi^{-1}_{N\,k} \circ \phi_{N\,k}) \circ f \circ (\phi^{-1}_{M\,i} \circ \phi_{M\,i} \circ \phi^{-1}_{M\,j}) \\ &= \varphi_{N,k,l} \circ (\phi_{N\,k} \circ f \circ \phi^{-1}_{M\,i}) \circ \varphi_{M,j,i} = \varphi_{N,k,l} \circ f_{i\,k} \circ \varphi_{M,j,i}, \end{aligned}$$

the definition does not depend on the which charts are used.

Example 2.13. We have already seen that open subsets of euclidean space are manifolds in a particularly simple way, namely the coordinates and parametrisations are the identity function. Therefore a function in charts is the same thing as a function. This shows that a function between euclidean spaces is C^ℓ according to the manifold definition if and only if it is C^ℓ according to the ordinary definition.

Example 2.14. Consider the sphere $\mathbb{S}^2$ and the function $f : \mathbb{S}^2 \to \mathbb{R}$ given by $f(p^1, p^2, p^3) = p^3$. This is the height function. We can look at this function in charts

$$f_{N\,1} : U_N \to \mathbb{R} \qquad f_{S\,1} : U_S \to \mathbb{R}$$
$$f_{N\,1}(x) = \frac{\|x\|^2 - 1}{\|x\|^2 + 1} \qquad f_{S\,1}(y) = \frac{1 - \|y\|^2}{\|y\|^2 + 1}.$$

It may help to understand this if we calculate it for a few points. Consider the south pole $S = (0, 0, -1)$, which has a height $f(0, 0, -1) = -1$. In ϕ_N-coordinates the south pole is $(0, 0) \in U_N$ and $f_{N\,1}(0, 0) = -1$ as expected. Now consider the point $p = (0.64, 0.10, 0.76) \in \mathbb{S}^2$, which corresponds to Mannheim. Clearly is has $f(p) = 0.76$. In ϕ_N-coordinates it is

[3] It is necessary to require that $\phi_i(p)$ has an open neighbourhood $U \subset U_{M\,i}$ such that $f[U] \subset U_{N\,j}$. If you omit this condition, it is possible to make an example where f is continuous in every chart but is not continuous on M as a whole.

$(2.7, 0.4) \in U_N$ and in ϕ_S-coordinates it is $(0.36, 0.06) \in U_S$. So then we compute

$$f_{N\,1}(2.7, 0.4) = \frac{7.3 - 1}{7.3 + 1} = 0.76, \qquad f_{S\,1}(0.36, 0.06) = \frac{1 - 0.135}{0.135 + 1} = 0.76.$$

In conclusion, f in charts is nothing other than a manipulation of the formula for f to use coordinates; it gives the same result.

Example 2.15. Consider Example 2.5, where we glued to intervals to make a circle. We can give a function f that *embeds* M into euclidean space $\mathbb{R}^2$. It is easier to write the formulas in charts. We define

$$\begin{aligned} f_{1\,1}(x) &= (\cos \pi x, \sin \pi x)\,, \\ f_{2\,1}(y) &= (\cos \pi(y+1), \sin \pi(y+1))\,. \end{aligned}$$

We can see that this is well-defined in the following way. Take a point $x \in \operatorname{dom} \varphi_{1\,2}$. Then

$$\begin{aligned} f_{2\,1}(\varphi_{1\,2}(x)) &= (\cos \pi(\varphi_{1\,2}(x) + 1), \sin \pi(\varphi_{1\,2}(x) + 1)) \\ &= \begin{cases} (\cos \pi(x + 1 + 1), \sin \pi(x + 1 + 1)) & \text{for } x \in (-1, 0) \\ (\cos \pi(x - 1 + 1), \sin \pi(x - 1 + 1)) & \text{for } x \in (0, 1) \end{cases} \\ &= \begin{cases} (\cos(\pi x + 2\pi), \sin(\pi x + 2\pi)) & \text{for } x \in (-1, 0) \\ (\cos \pi x, \sin \pi x) & \text{for } x \in (0, 1) \end{cases} \\ &= (\cos \pi x, \sin \pi x) \\ &= f_{1\,1}(x). \end{aligned}$$

What we have shown is that if $x \in U_1 \sim y \in U_2$ then $f_{1\,1}(x) = f_{2\,1}(y)$. This means that it doesn't matter which chart you use, the result is the same. In other words we have defined a function f that doesn't depend on charts, f is defined on M.

This example shows us how to embed Example 2.5 into $\mathbb{R}^2$ to get what we would normally think of as the circle. But please keep in mind that manifolds are defined as the gluing of charts; they are defined as an abstract space that does not need live in a bigger space. There are many ways to embed the circle into euclidean space. Even when we define a manifold starting with a subset of euclidean space, we leave the embedding behind.

2.3 Vectors

Our definition of manifolds makes it easy to define (tangent) vectors and vector fields. At any point $x \in U_1 \subset \mathbb{R}^n$ we have the tangent vectors $\{(v^1, \dots, v^n) \in \mathbb{R}^n\}$ and a vector field on U_1 is a function $X : U_1 \to \mathbb{R}^n$. To make this into a definition on a manifold, however, we need a way to make this independent of the chart. Alternatively, we need a way to compare vectors that are defined using different charts. There are essentially two equivalent ways to do this: using curves and using directional derivatives.

Both methods have the same setup. Let $U_1, U_2 \subset \mathbb{R}^n$ be two charts and $\varphi : U_1 \rightharpoonup U_2$ the transition (we leave off the subscripts for this explanation to simplify notation). Let x be a point in $\operatorname{dom}\varphi$, $y = \varphi(x)$, and $v = (v^1, \dots, v^n)$ be a vector on U_1 and $w = (w^1, \dots, w^n)$ be a vector on U_2.

Curve method: Consider the curve $\alpha(t) = x + vt \in U_1$. This curve has $\alpha(0) = x$ and $\alpha'(0) = v$. Using the transition, we also have a curve $\beta = \varphi \circ \alpha$ in U_2 with $\beta(0) = \varphi(x) = y$. The idea is that v in the first chart is transformed to $w = \beta'(0)$ in the second chart. Using the chain rule

$$w = \beta'(0) = (J_x\varphi)\,\alpha'(0) = (J_x\varphi)\,v,$$

where $J_x\varphi$ is the matrix of partial derivatives of φ, also called the Jacobian matrix, evaluated at the point x.

Example 2.16. Consider the plane $\mathbb{R}^2$ with cartesian coordinates (y^1, y^2), but also polar coordinates $(x^1, x^2) = (r, \theta)$. The transition function from polar to cartesian is $\varphi(r, \theta) = (r\cos\theta, r\sin\theta)$. For example, the point with $(r, \theta) = (1, \pi/4)$ is $(y^1, y^2) = (1/\sqrt{2}, 1/\sqrt{2})$. The Jacobian of φ at this point is

$$J_{(1,\pi/4)}\varphi = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}_{(1,\pi/4)} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

The forgoing calculation tells us that the radial vector, $(1, 0)$ in polar coordinates, is transformed to

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

in cartesian coordinates. Likewise the rotation vector, $(0, 1)$ in polar coordinates, corresponds to the vector $(-1/\sqrt{2}, 1/\sqrt{2})$ in cartesian coordinates. This is exactly as we would expect.

Directional derivative method: Take any function $h : U_2 \to \mathbb{R}$. We can use the transition function to write h as a function on U_1, at least near x, namely $h \circ \varphi$. We compute the derivative in the direction of v at the point x and apply the chain rule

$$\sum_{j=1}^{n} v^j \frac{\partial(h\circ\varphi)}{\partial x^j}(x) = \sum_{i,j=1}^{n} v^j \frac{\partial h}{\partial y^i}(\varphi(x))\frac{\partial y^i}{\partial x^j}(x) = \sum_{i,j=1}^{n} v^j \frac{\partial h}{\partial y^i}(y)(J_x\varphi)_{ij} = \sum_{i=1}^{n}(J_x\varphi\, v)_i \frac{\partial h}{\partial y^i}(y).$$

We see that the derivative of $h \circ \varphi$ at x in the direction of v is equal to the derivative of h at y in the direction $w = J_x\varphi\, v$. From both of these methods we get the same answer:

Definition 2.17. *The vector v at $x \in U_1$ is equivalent to the vector w at $y \in U_2$ when $w = J_x\varphi\, v$. This formula is called the change of coordinates for vectors. We denote the set of equivalence classes of vectors at a point $p \in M$ by T_pM, called the* tangent space *to M at p.*

Because of the second method, it is common to write a tangent vector as $\sum_{i=1}^n v^i \frac{\partial}{\partial x^i}$. This notation has the advantage that the change of coordinates is built into the notation via the chain rule, as above. In fact, the $\frac{\partial}{\partial x^i}$ are a basis of the vector fields. Every vector field on U_1 can be written as a function

$$X : x \mapsto \sum_{i=1}^n v^i(x) \frac{\partial}{\partial x^i}$$

for functions $v^i : U_1 \to \mathbb{R}$. Because vector fields thought of this way can be evaluated at a point to give a tangent vector as well as act on a function, the notation $X(x)$ is potentially ambiguous. We will use $X|_p$ for evaluation and $X(f)$ for action on a function.

Example 2.18. Let's see how the coordinate vector fields transform between the two charts of stereographic projection on $\mathbb{S}^2$. We have from Exercise 2.11 that $y = \varphi(x) = \|x\|^{-2}x$. Note the useful relation $\|y\| = \|x\|^{-2}\|x\| = \|x\|^{-1}$. The Jacobian of the transition function is

$$\begin{aligned} J\varphi = \begin{pmatrix} \frac{\partial y^1}{\partial x^1} & \frac{\partial y^1}{\partial x^2} \\ \frac{\partial y^2}{\partial x^1} & \frac{\partial y^2}{\partial x^2} \end{pmatrix} &= \frac{1}{\|x\|^4} \begin{pmatrix} (x^2)^2 - (x^1)^2 & -2x^1x^2 \\ -2x^1x^2 & (x^1)^2 - (x^2)^2 \end{pmatrix} \\ &= \begin{pmatrix} (y^2)^2 - (y^1)^2 & -2y^1y^2 \\ -2y^1y^2 & (y^1)^2 - (y^2)^2 \end{pmatrix}. \end{aligned}$$

From this information we have

$$\begin{aligned} \frac{\partial}{\partial x^1} &= \frac{\partial y^1}{\partial x^1}\frac{\partial}{\partial y^1} + \frac{\partial y^2}{\partial x^1}\frac{\partial}{\partial y^2} = \Big[(y^2)^2 - (y^1)^2\Big]\frac{\partial}{\partial y^1} + \Big[-2y^1y^2\Big]\frac{\partial}{\partial y^2}, \\ \frac{\partial}{\partial x^2} &= \frac{\partial y^1}{\partial x^2}\frac{\partial}{\partial y^1} + \frac{\partial y^2}{\partial x^2}\frac{\partial}{\partial y^2} = \Big[-2y^1y^2\Big]\frac{\partial}{\partial y^1} + \Big[(y^1)^2 - (y^2)^2\Big]\frac{\partial}{\partial y^2}. \end{aligned}$$

We should note that this equality, which is really equivalence of vectors as per Definition 2.17, holds on the overlap between the two charts, namely away from the north and south poles. In particular, the vector $\frac{\partial}{\partial x^1}$ is not defined at the north pole, this expression only has meaning on the chart U_N. If we were to define a vector field by the formula

$$X = \begin{cases} \frac{\partial}{\partial x^1} & \text{for } x \in U_N, \\ \Big[(y^2)^2 - (y^1)^2\Big]\frac{\partial}{\partial y^1} + \Big[-2y^1y^2\Big]\frac{\partial}{\partial y^2} & \text{for } y \in U_S \end{cases}$$

then this is a well-defined vector field on all of $\mathbb{S}^2$, because it gives a vector at every point and on overlaps the two cases give equivalent vectors. Observe that X has a zero at $y = 0$. It is a theorem, called humorously the *Hairy ball theorem* or the *Hedgehog theorem*, that every vector field on $\mathbb{S}^2$ must have at least one zero.

The main lesson of this example is that what looks like the 'same vector' at two different points in one chart look like completely different vectors in another chart. From the above example, $\frac{\partial}{\partial x_1}$ at $x = (1, 0)$ and $\frac{\partial}{\partial x_1}$ at $x = (0, 1)$ look the same in the U_N chart, but in the U_S chart these

two vectors are

$$-\left.\frac{\partial}{\partial y^1}\right|_{(1,0)} \qquad \text{and} \qquad \left.\frac{\partial}{\partial y^1}\right|_{(0,1)}$$

respectively. This is because the Jacobian $J_x\varphi$ has a dependence on x, on the point in the manifold. Hence the equivalence relation is different at different points. The consequence is that there is no easy way to identity tangent vectors at different points of a manifold and we must give up on this notion for now. We will examine how this problem can be overcome in the next chapter, which concerns 'connections' between different points of a manifold.

Example 2.19. Let us do another example on $\mathbb{S}^2$. Consider the vector field

$$X = \begin{cases} -x^2\frac{\partial}{\partial x^1} + x^1\frac{\partial}{\partial x^2} & \text{for } x \in U_N, \\ -y^2\frac{\partial}{\partial y^1} + y^1\frac{\partial}{\partial y^2} & \text{for } y \in U_S \end{cases}$$

We first check that this is a well-defined vector field. We need to show that the vectors of the two cases are equivalent at corresponding points. We calculate

$$\begin{aligned} -x^2\frac{\partial}{\partial x^1} + x^1\frac{\partial}{\partial x^2} &= -\|y\|^{-2}y^2\left(\Big[(y^2)^2 - (y^1)^2\Big]\frac{\partial}{\partial y^1} + \Big[-2y^1y^2\Big]\frac{\partial}{\partial y^2}\right) \\ &\qquad + \|y\|^{-2}y^1\left(\Big[-2y^1y^2\Big]\frac{\partial}{\partial y^1} + \Big[(y^1)^2 - (y^2)^2\Big]\frac{\partial}{\partial y^2}\right) \\ &= \|y\|^{-2}\Big[-(y^2)^3 + (y^1)^2y^2 - 2(y^1)^2y^2\Big]\frac{\partial}{\partial y^1} \\ &\qquad + \|y\|^{-2}\Big[2y^1(y^2)^2 + (y^1)^3 - y^1(y^2)^2\Big]\frac{\partial}{\partial y^2} \\ &= -y^2\frac{\partial}{\partial y^1} + y^1\frac{\partial}{\partial y^2}, \end{aligned}$$

So indeed they are equivalent.

We can also apply this vector field to the height function f from Example 2.14. In the chart U_N we have

$$Xf = -x^2\frac{\partial f_{N\,1}}{\partial x^1} + x^1\frac{\partial f_{N\,1}}{\partial x^2} = -x^2\frac{4x^1}{((x^1)^2 + (x^2)^2 + 1)^2} + x^1\frac{4x^2}{((x^1)^2 + (x^2)^2 + 1)^2} = 0.$$

Likewise in the chart U_S we have

$$Xf = -y^2\frac{\partial f_{S\,1}}{\partial y^1} + y^1\frac{\partial f_{S\,1}}{\partial y^2} = 0.$$

One way to understand this result is to look at the vector fields and the height function in charts. The vector field 'goes' in a circle, it points along the circles $\|x\| = \text{const}$. The height function f is constant along these circles. So naturally Xf, which is the change in f in the direction of the vector field X, is zero.

Example 2.20. We consider the manifold from Example 2.5, the circle constructed from two intervals. Because the Jacobian of the transition function is just the constant 1, the following

is a well-defined vector field:

$$X = \begin{cases} \frac{\partial}{\partial x} & \text{for } x \in U_1, \\ \frac{\partial}{\partial y} & \text{for } y \in U_2 \end{cases}$$

We can act this vector field on the function $f : \mathbb{S}^1 \to \mathbb{R}$ defined in charts as

$$f_{1\,1}(x) = \cos \pi x, \qquad f_{2\,1}(y) = \cos \pi(y+1),$$

which is just the first component of the function from Example 2.15. We get

$$Xf_{1\,1}(x) = \frac{\partial}{\partial x} \cos \pi x = -\pi \sin \pi x, \qquad Xf_{2\,1}(y) = \frac{\partial}{\partial y} \cos \pi(y+1) = -\pi \sin \pi(y+1).$$

This is a well-defined function on $\mathbb{S}^1$, which we know because it is $-\pi$ multiplied by the second component from Example 2.15.

This last example illustrates that a vector field X applied in a chart to a function f_i is another function Xf_i. The fact that we used the chain rule to define equivalence of vectors is precisely the condition to ensure that the resulting functions in charts piece together to give a well-defined function Xf on the whole manifold.

If we have a function between two manifolds $f : M \to N$, the *tangent map* or *pushforward* of f at p is a function between the tangent spaces $T_pf : T_pM \to T_pN$. We will define it in two ways. The easiest definition is using the curve method of vectors. If we have a vector $v \in T_xM$ then in a chart U_i the curve $\alpha(t) = x + vt$ is a representative of this vector. Observe that $f \circ \alpha : (a, b) \to N$ is a curve in N. We define $T_pf(v)$ to be the tangent vector of the curve $f \circ \alpha$ at $t = 0$.

We also give a practical formula for calculating the tangent map. The curve $f \circ \alpha$ must lie in some chart V_j of N. In charts we have $\alpha_i : (a, b) \to U_i$ and $f_{i\,j} : U_i \rightharpoonup V_j$. So the tangent vector of the curve is

$$\frac{d}{dt} f_{i\,j}(\alpha(t))\Big|_{t=0} = (J_{\alpha(t)} f_{i\,j})\, \alpha'(t)\Big|_{t=0} = (J_{\alpha(0)} f_{i\,j})\, \alpha'(0) = (J_x f_{i\,j})\, v,$$

using the chain rule and the fact that $\alpha(0) = x$, $\alpha'(0) = v$. Thus we see in charts that the tangent map is the Jacobian of $f_{i\,j}$.

We could have also given a definition of the pushforward based on the directional derivative idea. If $w = T_pf(v) \in T_{f(p)}N$ is the pushforward of a vector $v \in T_pM$, then we can ask how v acts on a function $g : N \to \mathbb{R}$. The observation is that $g \circ f : M \to \mathbb{R}$ and

$$w(g) = T_pf(v)(g) = v(g \circ f).$$

Example 2.21. Consider the vector field X from Example 2.19 and inverse stereographic projection from Example 2.10. In fact, the functions in that example together give a function $\phi^{-1} : \mathbb{S}^2 \to \mathbb{R}^3$. We can ask what the pushforward of a vector $X|_p \in T_p\mathbb{S}^2$ to $T_{\phi^{-1}(p)}\mathbb{R}^3$ is. For simplicity, assume that p is not the north pole, so we can just work with ϕ_N^{-1}. We need the

Jacobian:

$$\phi_N^{-1} = \frac{1}{(x^1)^2+(x^2)^2+1}\left(2x^1, 2x^2, (x^1)^2+(x^2)^2-1\right)$$
$$J_x(\phi_N^{-1}) = \frac{2}{(\|x\|^2+1)^2}\begin{pmatrix} -(x^1)^2+(x^2)^2+1 & -2x^1x^2 \\ -2x^1x^2 & (x^1)^2-(x^2)^2+1 \\ 2x^1 & 2x^2 \end{pmatrix}.$$

The point is that in the coordinates vector basis on U_N, the vector field X is $(-x^2, x^1)$. So any vector of this form is sent by the tangent map of ϕ^{-1} to

$$\begin{aligned} J_x(\phi_N^{-1})X|_x &= \frac{2}{(\|x\|^2+1)^2}\begin{pmatrix} -(x^1)^2+(x^2)^2+1 & -2x^1x^2 \\ -2x^1x^2 & (x^1)^2-(x^2)^2+1 \\ 2x^1 & 2x^2 \end{pmatrix}\begin{pmatrix} -x^2 \\ x^1 \end{pmatrix} \\ &= \frac{2}{(\|x\|^2+1)^2}\begin{pmatrix} (-(x^1)^2+(x^2)^2+1)(-x^2)-2(x^1)^2x^2 \\ 2x^1(x^2)^2+((x^1)^2-(x^2)^2+1)x^1 \\ -2x^1x^2+2x^1x^2 \end{pmatrix} \\ &= \frac{2}{(\|x\|^2+1)^2}\begin{pmatrix} ((x^1)^2+(x^2)^2+1)(-x^2) \\ ((x^1)^2+(x^2)^2+1)x^1 \\ 0 \end{pmatrix} \\ &= \frac{2}{\|x\|^2+1}\begin{pmatrix} -x^2 \\ x^1 \\ 0 \end{pmatrix}. \end{aligned}$$

This give an alternative way to understand Example 2.19. We see that, viewed in $\mathbb{R}^3$, the vectors of this field are always tangent to a line of latitude. It is no wonder therefore that the derivative of the height function in the directions of these vectors is zero.

An important observation should be made about the previous example. We were careful to stress that the tangent map of ϕ^{-1} was applied to vectors and not a vector field, even though the vectors in this case came from a vector field. That is because the tangent map does not, as students sometimes assume, take vector fields to vector fields. There are two ways it fails to do this. In the above example we see that each vector field X is transformed to a vector on $\mathbb{R}^3$, but not every point of $\mathbb{R}^3$ gets a vector. Only the points in the image of ϕ^{-1} get a vector. Therefore the result is not a vector field on $\mathbb{R}^3$.

The other way that the tangent map can fail to transform a vector field into another vector field, which does not apply to the above example, is if the function is not injective. In that case some points in the image get multiple vectors from the tangent map. Since a vector field is a certain type of function, and functions must have exactly one output, the pushforward of a vector field may not be a function.

We end this section by highlighting a distinction that we have elided until now. A tangent vector is an intrinsic concept on a manifold, it does not require the manifold to be immersed in euclidean space. And yet in the previous chapter we repeatedly talked about tangent vectors to surfaces as a vector in $\mathbb{R}^3$ that was tangent in the geometrical sense. The connections between these two ideas is exactly the idea of immersion and pushforward.

Definition 2.22. *A function $f : M \to N$ is an* immersion *at $p \in M$ if T_pf is injective.*

In Example 2.15 we saw that a manifold can be mapped into a euclidean space by a function. The pushforward of its tangent vectors by this function are exactly the tangent vectors in the intuitive picture. For another example consider again Example 2.21. Here we took the vector $X|_x = (-x^2, x^1)$ in the U_N chart and pushed it into euclidean space $\mathbb{R}^3$ to get a vector that is tangent to the unit sphere. Thus for an immersion, it is possible to consider the elements of T_pM as certain elements of T_pN, and this is unambiguous because by definition T_pf is injective.

2.4 Vector Bundles

This section will not be used later in the script. It exists just to flex how natural the construction of the tangent bundle is in our approach to manifolds.

Above we have defined the tangent space T_pM to a manifold M at a point $p \in M$ and discussed how the vectors of T_pM and T_qM should be thought of as distinct, even if p and q are in the same chart. But we also worked with examples of vector fields, whose values are vectors at different points. We can reconcile this tension by putting all the vectors of a manifold together into a new manifold:

Definition 2.23. *Let M be a manifold with atlas $\mathcal{A} = (n, \mathcal{I}, \{U_i\}_{i\in\mathcal{I}}, \{\varphi_{ij}\}_{i,j\in\mathcal{I}})$. We construct a new atlas $T\mathcal{A} = (2n, \mathcal{I}, \{TU_i\}_{i\in\mathcal{I}}, \{T\varphi_{ij}\}_{i,j\in\mathcal{I}})$, where $TU_i = U_i \times \mathbb{R}^n$ and*

$$T\varphi_{ij} : \operatorname{dom}\varphi_{ij} \times \mathbb{R}^n \to \operatorname{img}\varphi_{ij} \times \mathbb{R}^n, \qquad (x, v) \mapsto (\varphi_{ij}(x), (J_x\varphi_{ij})v).$$

The corresponding manifold is called the tangent bundle *TM. There is a function $\pi : TM \to M$ in charts by*

$$\pi_{ii}(x, v) = x.$$

We tend to think of the tangent bundle of M as all the vectors of M, with the understanding that vectors at different points are distinct from one another. The function π is called the canonical projection of the bundle. Intuitively it takes a tangent vector to its base point. Hence $\pi^{-1}[\{p\}]$ are all the vectors which live at the same point, in other words T_pM.

The tangent bundle allows us to speak and reason formally about tangent vectors as a whole. We can define a vector field on M as a function X from M to TM with the property that $\pi \circ X = \mathrm{id}_M$. This property says that for every point $p \in M$ the vector $X|_p$ must have the base point $\pi(X|_p) = p$, which is exactly what a vector field is.

Exercise 2.24. Check that the tangent bundle is a manifold. In particular you must check the unique limit condition.

Example 2.25. Since the trivial atlas for an open subset $U \subset \mathbb{R}^n$ has only one chart $U_0 = U$, so too does its tangent bundle $TU_0 = U \times \mathbb{R}^n$. Therefore the tangent bundle is just the cartesian product. The only difference between TU and usual way we think about vectors in euclidean space is that the base point of the vector is important for the tangent bundle.

Example 2.26. We saw in Example 2.20 that there was a vector field X on $\mathbb{S}^1$ that was never zero. Because $T_p\mathbb{S}^1$ is one dimensional, every vector in it must be a scalar multiple of $X|_p$. This means we can define a function from T_pM to $\mathbb{R}$

$$v \mapsto a, \text{ where } v = aX|_p.$$

Allowing p to change gives us a function from $T\mathbb{S}^1$ to $\mathbb{S}^1 \times \mathbb{R}$. Thus the tangent bundle of the circle is also a product. Tangent bundles that a products are called *trivial*.

Example 2.27. In this example we want to show a non-trivial tangent bundle, but it's actually somewhat difficult to prove non-triviality. Instead we will try to convey the idea. The tangent bundle of $\mathbb{S}^2$ is non-trivial. If it were trivial, that would mean there would exist a smooth bijective correspondence between $T\mathbb{S}^2$ and $\mathbb{S}^2 \times \mathbb{R}^2$. We could use this correspondence to write any vector field X on $\mathbb{S}^2$ as $(p, \tilde{X}(p))$ with $\tilde{X}(p) \in \mathbb{R}^2$. Conversely there would exist a vector field with the form $(p, (1, 0))$. This vector field is never zero. We have already mentioned (but not proved) the hairy ball theorem: every vector field on the sphere has at least one zero. This is one reason that the tangent bundle of $\mathbb{S}^2$ cannot be trivial.

The tangent map takes its nicest form when expressed with the tangent bundle. We can collect together all the tangent maps $T_p f : T_p M \to T_p N$ into a single map $Tf : TM \to TN$. Using the formula in charts in terms of the Jacobian we get

$$(Tf)_{i\,j}(x, v) = \Big(f_{i\,j}(x), (J_x f_{i\,j})\, v\Big) \in V_j \times \mathbb{R}^n = TV_j.$$

This formula shows that the tangent map is the generalisation of the Jacobian, and all local properties of the Jacobian carry over to the tangent map.

Exercise 2.28. Prove that the above formula for the tangent map is well-defined (in different charts) by using the chain rule for Jacobians: for $y = \varphi_{M,i,j}(x)$ and $z = f_{j\,k}(y)$

$$J_x(\varphi_{N,k,l} \circ f_{j\,k} \circ \varphi_{M,i,j}) = (J_z \varphi_{N,k,l})(J_y f_{j\,k})(J_x \varphi_{M,i,j}).$$

Exercise 2.29. Calculate the tangent map for f from Example 2.15. Use this to give an alternative description of the tangent bundle $T\mathbb{S}^1$ by locating it as a set inside $T\mathbb{R}^2 = \mathbb{R}^2 \times \mathbb{R}^2$.

2.5 Summation Convention

As you have seen, when working with vectors in charts, there are many Σ summations, but all of them are just from $i = 1$ to n. There is a convention, called the Einstein summation convention or the Einstein rule, that allows us to omit all these sigmas.

Definition 2.30 (Einstein Summation)**.** *We apply the following notational rule: when an index occurs twice in a term, once as an upper index and once as a lower index, we sum that index over its range. For the purposes of this rule $\frac{\partial}{\partial x^i} = \partial_i$ counts as a lower index.*

In this notation, the chain rule reads

$$\frac{\partial}{\partial x^1} = \frac{\partial y^1}{\partial x^1}\frac{\partial}{\partial y^1} + \frac{\partial y^2}{\partial x^1}\frac{\partial}{\partial y^2} = \sum_{i=1}^{2} \frac{\partial y^i}{\partial x^1}\frac{\partial}{\partial y^i} = \frac{\partial y^i}{\partial x^1}\frac{\partial}{\partial y^i}.$$

A vector field X that has coefficients X^i with respect to a chart can neatly be written as $X = X^i\partial_i$, and its expression with respect to a second chart y might be written using the equivalences of vectors as

$$X = X^i\frac{\partial}{\partial x^i} = X^i\frac{\partial y^j}{\partial x^i}\frac{\partial}{\partial y^j}.$$

This convention is useful for even ordinary linear algebra. If you have a matrix A we would normally write the entries A_{ij}. However if we instead write it A^i_j with a mix of upper and lower indices then we can write matrix multiplication in the following way

$$Av = \left(A^i_j\right)\left(v^j\right) = \sum_{j=1}^{n} A^i_j v^j = A^i_j v^j.$$

Likewise matrix multiplication AB would be $(AB)^i_k = A^i_j B^j_k$. The saying to turn matrix algebra into index algebra with summation convention is "Upper indices go up to down; lower indices go left to right".

We know that bilinear functions can also be represented as a matrix; in usual matrix notation:

$$g(v, w) = v^T\left(g_{ij}\right)w = \sum_{i,j=1}^{n} v^i g_{ij} w^j.$$

In summation convention we write this as $g_{ij}v^i w^j$, with the vector components having upper indices, which forces g to have two lower indices. This is actually an advantage of index notation over matrix notation. Though linear transforms and bilinear transforms can both be represented as a matrix, those matrix representations behave differently under change of basis. We can see that they are different in index notation, but not in matrix notation.

Some authors do not have the condition that you need one upper and one lower index, and allow summation over any repeated index. I do not like this, because it makes it impossible to talk about the diagonal elements g_{ii}. If you need to sum over two indices on the same level, you can of course use a summation sign. Alternatively you can use an identity matrix, whose entries are usually called δ_{ij}. For example

$$\operatorname{tr}\left(g_{ij}\right) = \sum_{i=1}^{n} g_{ii} = \delta^{i,j} g_{ij}.$$

Finally, some useful rules of calculation are

$$\frac{\partial x^i}{\partial y^j}\frac{\partial y^j}{\partial x^k} = \delta^i_k \qquad \text{and} \qquad v^i\delta^j_i = v^j.$$

2.6 The Lie Bracket

We have seen examples of how vectors are a type of directional derivative of functions and that a vector field applied to a function gives another function. A natural question is to ask whether we can also differentiate a vector field using a vector field. The answer is yes: in fact vector fields can differentiate many objects on manifolds using a process called the Lie derivative $\mathcal{L}_X$, but the process complicated and we won't explain it here. We show with the following example why the 'obvious' way to differentiate vector fields doesn't work.

Example 2.31. Recall Example 2.18 on $\mathbb{S}^2$ of

$$X = \begin{cases} \frac{\partial}{\partial x^1} & \text{for } x \in U_N, \\ \Big[(y^2)^2 - (y^1)^2\Big]\frac{\partial}{\partial y^1} + \Big[-2y^1y^2\Big]\frac{\partial}{\partial y^2} & \text{for } y \in U_S \end{cases}$$

You might guess that you can differentiate a vector field $X = X^i\frac{\partial}{\partial x^i}$ just by differentiating its coefficient functions X^i,

$$Y(X) \overset{?}{:=} \sum_{i=1}^{n} Y(X^i)\frac{\partial}{\partial x^i},$$

but this doesn't produce a well-defined vector field, because in different charts it produces inequivalent vectors. For example, if we act X on X with this definition, in U_N we have

$$\left[\frac{\partial}{\partial x^1}1\right]\frac{\partial}{\partial x^1} + \left[\frac{\partial}{\partial x^1}0\right]\frac{\partial}{\partial x^2} = 0,$$

but in U_S we have

$$\begin{aligned} &X\Big((y^2)^2 - (y^1)^2\Big)\frac{\partial}{\partial y^1} + X\Big(-2y^1y^2\Big)\frac{\partial}{\partial y^2} \\ &= \Big[((y^2)^2 - (y^1)^2)(-2y^1) + (-2y^1y^2)(2y^2)\Big]\frac{\partial}{\partial y^1} \\ &\quad + \Big[((y^2)^2 - (y^1)^2)(-2y^2) + (-2y^1y^2)(-2y^1)\Big]\frac{\partial}{\partial y^2} \neq 0. \end{aligned}$$

It turns out the correct way to differentiate vector fields is to take the commutator of the above guess.

Definition 2.32. *The Lie derivative of a vector field Y with respect to a vector field X, also called the Lie bracket, acts on a function f by*

$$(\mathcal{L}_X Y)(f) = [X, Y](f) := X(Y(f)) - Y(X(f)).$$

From this definition we can ask how to write it in charts:

$$\begin{aligned}
[X,Y](f) &= X(Y(f)) - Y(X(f)) = X\left(Y^j\frac{\partial f}{\partial x^j}\right) - Y\left(X^i\frac{\partial f}{\partial x^i}\right)\\
&= X^i\frac{\partial}{\partial x^i}\left(Y^j\frac{\partial f}{\partial x^j}\right) - Y^j\frac{\partial}{\partial x^j}\left(X^i\frac{\partial f}{\partial x^i}\right)\\
&= X^i\frac{\partial Y^j}{\partial x^i}\frac{\partial f}{\partial x^j} + X^iY^j\frac{\partial^2 f}{\partial x^i\partial x^j} - Y^j\frac{\partial X^i}{\partial x^j}\frac{\partial f}{\partial x^i} - Y^jX^i\frac{\partial^2 f}{\partial x^j\partial x^i}\\
&= \left[X^j\frac{\partial Y^i}{\partial x^j} - Y^j\frac{\partial X^i}{\partial x^j}\right]\frac{\partial f}{\partial x^i} = \left[X(Y^i) - Y(X^i)\right]\frac{\partial f}{\partial x^i}.
\end{aligned}$$

So even though the definition of $[X,Y]$ appears to have second derivatives of f, these cancel out and what remains is indeed a (first order) vector field acting on f.

Whenever we have an expression in charts, it is good to check that in different charts it produces equivalent vectors. We do so in the following lemma. This is not strictly necessary, because the $[X,Y](f) = X(Y(f)) - Y(X(f))$ definition doesn't use charts, but good practice none-the-less.

Lemma 2.33. *The above chart expression gives a vector field that is independent of the choice of chart.*

Proof. If $X = X^i(x)\frac{\partial}{\partial x^i}$ and $Y = Y^i(x)\frac{\partial}{\partial x^i}$ are the expressions of two vector fields in the x-chart, then the expression of these fields in the y-chart is

$$X = X^j\frac{\partial y^i}{\partial x^j}\frac{\partial}{\partial y^i}, \qquad Y = Y^j\frac{\partial y^i}{\partial x^j}\frac{\partial}{\partial y^i}.$$

In this chart, the ith component of $[X,Y]$ is

$$\begin{aligned}
X(\tilde{Y}^i) - Y(\tilde{X}^i) &= X^l\frac{\partial y^k}{\partial x^l}\frac{\partial}{\partial y^k}\left(Y^j\frac{\partial y^i}{\partial x^j}\right) - Y^j\frac{\partial y^k}{\partial x^j}\frac{\partial}{\partial y^k}\left(X^l\frac{\partial y^i}{\partial x^l}\right)\\
&= X^l\frac{\partial y^k}{\partial x^l}\frac{\partial}{\partial y^k}\left(Y^j\frac{\partial y^i}{\partial x^j}\right) - Y^j\frac{\partial y^k}{\partial x^j}\frac{\partial}{\partial y^k}\left(X^l\frac{\partial y^i}{\partial x^l}\right)\\
&= X^l\frac{\partial}{\partial x^l}\left(Y^j\frac{\partial y^i}{\partial x^j}\right) - Y^j\frac{\partial}{\partial x^j}\left(X^l\frac{\partial y^i}{\partial x^l}\right)\\
&= X^l\frac{\partial Y^j}{\partial x^l}\frac{\partial y^i}{\partial x^j} + X^lY^j\frac{\partial^n y^i}{\partial x^l\partial x^j} - Y^j\frac{\partial X^l}{\partial x^j}\frac{\partial y^i}{\partial x^l} - Y^jX^l\frac{\partial^n y^i}{\partial x^j\partial x^l}\\
&= X^l\frac{\partial Y^j}{\partial x^l}\frac{\partial y^i}{\partial x^j} - Y^l\frac{\partial X^j}{\partial x^l}\frac{\partial y^i}{\partial x^j} = \left[X^l\frac{\partial Y^j}{\partial x^l} - Y^l\frac{\partial X^j}{\partial x^l}\right]\frac{\partial y^i}{\partial x^j}.
\end{aligned}$$

In total then, in the vector field $[X,Y]$ in the y-chart is

$$[X,Y] = \left[X^l\frac{\partial Y^j}{\partial x^l} - Y^l\frac{\partial X^j}{\partial x^l}\right]\frac{\partial y^i}{\partial x^j}\frac{\partial}{\partial y^i}.$$

But we see that this is equivalent to the expression in the x-chart. Therefore the definition produces a well-defined vector field. □

There are of course many things that can be said about the Lie bracket. The first observation is that it is $\mathbb{R}$-bilinear and antisymmetric: $[aX + b\tilde{X}, Y] = a[X,Y] + b[\tilde{X},Y]$ and $[X,Y] = -[Y,X]$.

If you have two coordinate vector fields, then their Lie bracket is zero. Thus one interpretation of the Lie bracket is that it is a measurement of how far two vector fields are from being coordinate vector fields. The final property that we will give is a product rule: if f is a function, then

$$[X, fY] = X(f)Y + f[X, Y],$$
$$[fX, Y] = -[Y, fX] = -Y(f)X - f[Y, X] = f[X, Y] - Y(f)X.$$

Example 2.34. We give some euclidean examples. Consider the plane $\mathbb{R}^2$. Let $X = \partial_1$ and $Y = \partial_2$. Then plugging in the definitions

$$[X, Y] = X(Y^i)\partial_i - Y(X^i)\partial_i = X(1)\partial_2 - Y(1)\partial_1 = 0.$$

Next consider $V = (1 + x^2)\partial_1$. Then

$$[V, Y] = V(1)\partial_2 - Y(1 + x^2)\partial_1 = -\partial_1.$$

Finally, set $W = x^2\partial_1 - x^1\partial_2$. then

$$[V, W] = V(x^2)\partial_1 - V(x^1)\partial_2 - W(1+x^2)\partial_1 = 0 - (1+x^2)\partial_2 - (0-x^1)\partial_1 = -(1+x^2)\partial_2 + x^1\partial_1.$$

Example 2.35. Consider again the sphere $\mathbb{S}^2$ and lets take the Lie bracket of

$$X = \begin{cases} \frac{\partial}{\partial x^1} & \text{for } x \in U_N, \\ \Big[(y^2)^2 - (y^1)^2\Big]\frac{\partial}{\partial y^1} + \Big[-2y^1y^2\Big]\frac{\partial}{\partial y^2} & \text{for } y \in U_S \end{cases}$$

and the latitude vector field

$$Y = \begin{cases} -x^2\frac{\partial}{\partial x^1} + x^1\frac{\partial}{\partial x^2} & \text{for } x \in U_N, \\ -y^2\frac{\partial}{\partial y^1} + y^1\frac{\partial}{\partial y^2} & \text{for } y \in U_S \end{cases}$$

In U_N the calculation is rather easy

$$[X, Y] = \Big[0 - 0\Big]\frac{\partial}{\partial x^1} + \Big[1 - 0\Big]\frac{\partial}{\partial x^2} = \frac{\partial}{\partial x^2}$$

In U_S the calculation is a little messier

$$\begin{aligned}[X, Y] &= \Big[((y^2)^2 - (y^1)^2)0 + (-2y^1y^2)(-1) - (-y^2)(-2y^1) - y^1(2y^2)\Big]\frac{\partial}{\partial y^1} \\ &\quad + \Big[((y^2)^2 - (y^1)^2)1 + (-2y^1y^2)0 - (-y^2)(-2y^2) - y^1(-2y^1)\Big]\frac{\partial}{\partial y^2} \\ &= \Big[-2y^1y^2\Big]\frac{\partial}{\partial y^1} + \Big[(y^1)^2 - (y^2)^2\Big]\frac{\partial}{\partial y^2}\end{aligned}$$

We know that these are equivalent on the overlap from Example 2.18.

Chapter 3

Metrics and Connections

3.1 Riemannian Metrics

Finally we come to the definition of a Riemannian metric, the object that gives this field its name. Let us dispel a common misunderstanding: a Riemannian metric is not a distance function, which goes against modern terminology (a la metric spaces). Instead it is a generalisation of an inner product. As we saw for surfaces, an inner product allows us to define a notion of length, so there is a close relation between distance functions and inner products on manifolds. But a new student to the field must get used to the change in terminology.

Definition 3.1. *A* Riemannian metric g *on a manifold* M *is a choice of inner product for every tangent space* T_pM*. If* U *is a chart of* M*, then we can express* g *in charts using the coordinate basis vectors:*

$$g_{ij}(p) = g\left(\left.\frac{\partial}{\partial x^i}\right|_p, \left.\frac{\partial}{\partial x^j}\right|_p\right).$$

A Riemannian metric should be smooth in the sense that the functions g_{ij} *are smooth in any chart. A manifold with a Riemannian metric is called a* Riemannian manifold*. Length of and angle between vectors* $X, Y \in T_pM$ *is defined in the usual way*

$$\|X\|_g := \sqrt{g(X,X)}, \qquad \cos\theta = \frac{g(X,Y)}{\|X\|\,\|Y\|}$$

The functions g_{ij} are sufficient to determine the inner product of any two vectors by bilinearity:

$$g\left(X^i\frac{\partial}{\partial x^i}, Y^j\frac{\partial}{\partial x^j}\right) = X^iY^jg\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = X^iY^jg_{ij}.$$

The symmetry and positive definiteness of g imply that the matrix (g_{ij}) is symmetric and positive definite.

Example 3.2. We have seen in Example 2.3 that any open subset of euclidean space is a manifold with one chart. It is also a Riemannian manifold with the usual dot product

$$g_{ij} = g\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = \frac{\partial}{\partial x^i}\cdot\frac{\partial}{\partial x^j} = \delta_{ij}.$$

Notice that the matrix of the metric in charts is symmetric and positive definite. This is also called the standard metric on $\mathbb{R}^n$.

Example 3.3. In fact we have seen Riemannian metrics already, namely the first fundamental form of a surface. For the helicoid, in Example 1.18, in the chart $U = \mathbb{R}^2$ we had coordinates $x^1 = u, x^2 = v$ and

$$g_{1\,1}(u,v) = 1, \qquad g_{1\,2}(u,v) = g_{2\,1}(u,v) = 0, \qquad g_{2\,2}(u,v) = u^2 + b^2.$$

For this example we see that $g_{i\,j}$ are non-constant functions (at least, $g_{2\,2}$ is non-constant). We understand that the length coordinate basis vector

$$\left\| \left. \frac{\partial}{\partial v} \right|_{(u,v)} \right\| = \sqrt{u^2 + b^2}$$

is different at different points of the helicoid.

We can ask how the functions $g_{i\,j}$ in a chart U are related to those $\tilde{g}_{i\,j}$ in an overlapping chart $\tilde{U}$. We know that the inner product should be independent of basis, so we compute it in two ways:

$$\tilde{g}_{i\,j} = g\left(\frac{\partial}{\partial y^i}, \frac{\partial}{\partial y^j}\right) = g\left(\frac{\partial x^k}{\partial y^i}\frac{\partial}{\partial x^k}, \frac{\partial x^l}{\partial y^j}\frac{\partial}{\partial x^l}\right) = \frac{\partial x^k}{\partial y^i}\frac{\partial x^l}{\partial y^j} g\left(\frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}\right) = \frac{\partial x^k}{\partial y^i}\frac{\partial x^l}{\partial y^j} g_{k\,l}.$$

Notice the subtle contrast to the equivalence relation for vectors:

$$v^i \frac{\partial}{\partial x^i} = v^i \frac{\partial y^j}{\partial x^i}\frac{\partial}{\partial y^j} = \tilde{v}^j \frac{\partial}{\partial y^j} \qquad \Rightarrow \qquad \tilde{v}^j = v^i \frac{\partial y^j}{\partial x^i}.$$

The term for objects that transform with $\frac{\partial x^k}{\partial y^i}$, like $g_{i\,j}$, is *covariant*, whereas those that transform with $\frac{\partial y^j}{\partial x^i}$, like the coefficients of vectors, are called *contravariant*. The convention is to use lower indices for covariant things, and upper indices for contravariant things. Historically this convention came before the summation convention. Because $\frac{\partial x^i}{\partial y^j}\frac{\partial y^j}{\partial x^i} = 1$ by the chain rule, when covariant and contravariant objects are 'multiplied', as in the above formula for g, then the result is independent of charts. This explains why there are so many sums of upper index with lower index, and was the motivation of the summation convention.

Clearly one can endow a manifold with functions $g_{i\,j}$ that satisfy the necessary properties and thereby make it a Riemannian manifold. But this is not usually how we construct Riemannian manifolds. It is far more common to 'inherit' a metric from a bigger Riemannian manifold. This is how we got a metric on the helicoid. In general, we use the tangent map to move vectors on one manifold into the tangent space of another.

Definition 3.4. *Let M be a manifold, N a Riemannian manifold with metric g. Let $f : M \to N$ be an* immersion. *That means that $T_p f$ is injective at every point. Then we define a metric f^*g on M, called the* pullback metric *or the* induced metric, *by*

$$f^*g(v,w) := g\left(T_p f(v), T_p f(w)\right)$$

for any $v, w \in T_p M$.

Exercise 3.5. The formula for f^*g is well-defined for all smooth functions $f : M \to N$, so why is it necessary that f is an immersion?

Let's go through how the definitions of Section 1.4 fit with the definitions in this section. First we have the definition of a regular parameterised surface $\Phi : U \to \mathbb{R}^3$, Definition 1.16. Φ is a function between euclidean spaces, so the tangent map is just the Jacobian $T_p\Phi = J_p\Phi$. The condition that the Jacobian is rank two is equivalent to it being injective by the rank-nullity theorem of linear algebra. Therefore regular and immersed are equivalent.

The first fundamental form is exactly the standard metric on $\mathbb{R}^3$ pullbacked by Φ. In the coordinate basis vectors, we have

$$\begin{aligned} g_{i\,j} &= \Phi^* g^{\mathbb{R}^3}\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = g^{\mathbb{R}^3}\left((J_p\Phi)^k_i \frac{\partial}{\partial x^k}, (J_p\Phi)^l_j \frac{\partial}{\partial x^l}\right) = (J_p\Phi)^k_i (J_p\Phi)^l_j g^{\mathbb{R}^3}\left(\frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}\right) \\ &= (J_p\Phi)^k_i (J_p\Phi)^l_j \delta_{k\,l} = \frac{\partial \Phi^k}{\partial x^i}\frac{\partial \Phi^l}{\partial x^j}\delta_{k\,l} = \frac{\partial \Phi}{\partial x^i}\cdot\frac{\partial \Phi}{\partial x^j}, \end{aligned}$$

which is the definition of the first fundamental form.

Example 3.6. Consider $\mathbb{S}^2$. What does the induced metric from $\mathbb{R}^3$ look like in stereographic coordinates? Well, we need to compute the pushforward of the coordinates vector fields and take the dot product. The pushforward was already computed for the U_N chart in Example 2.21:

$$\begin{aligned} J_x(\phi_N^{-1})\frac{\partial}{\partial x^1} &= \frac{2}{(\|x\|^2+1)^2}\begin{pmatrix} -(x^1)^2+(x^2)^2+1 & -2x^1x^2 \\ -2x^1x^2 & (x^1)^2-(x^2)^2+1 \\ 2x^1 & 2x^2 \end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix} \\ &= \frac{2}{(\|x\|^2+1)^2}\begin{pmatrix} -(x^1)^2+(x^2)^2+1 \\ -2x^1x^2 \\ 2x^1 \end{pmatrix}, \\ J_x(\phi_N^{-1})\frac{\partial}{\partial x^2} &= \frac{2}{(\|x\|^2+1)^2}\begin{pmatrix} -2x^1x^2 \\ (x^1)^2-(x^2)^2+1 \\ 2x^2 \end{pmatrix}. \end{aligned}$$

Therefore

$$\begin{aligned} g_{1\,1} &= \frac{4}{(\|x\|^2+1)^4}\left[(-(x^1)^2+(x^2)^2+1)^2+(-2x^1x^2)^2+(2x^1)^2\right] \\ &= \frac{4}{(\|x\|^2+1)^4}\left[(x^1)^4+(x^2)^4+1+2(x^1)^2(x^2)^2+2(x^1)^2+2(x^2)^2\right] \\ &= \frac{4}{(\|x\|^2+1)^4}\left[(x^1)^2+(x^2)^2+1\right] = \frac{4}{(\|x\|^2+1)^2}, \end{aligned}$$

similarly

$$\begin{aligned} &g_{1\,2} = g_{2\,1} \\ &= \frac{4}{(\|x\|^2+1)^4}\left[(-(x^1)^2+(x^2)^2+1)(-2x^1x^2)+(-2x^1x^2)((x^1)^2-(x^2)^2+1)+4x^1x^2\right] = 0, \end{aligned}$$

and

$$g_{2\,2} = \frac{4}{(\|x\|^2+1)^4}\left[(-2x^1x^2)^2 + ((x^1)^2 - (x^2)^2 + 1)^2 + (2x^1)^2\right] = \frac{4}{(\|x\|^2+1)^2}.$$

Exercise 3.7. Compute $\tilde{g}_{i\,j}$ in the chart U_S and verify the change of chart formula for the metric.

Finally, consider the notion of isometry in Definition 1.35. It says that two parameterised surfaces are isometric if their parametrisations induce equal metrics. We give the following more general definition.

Definition 3.8. *Let M, N be Riemannian manifolds and let $f : M \to N$ be an immersion. We call f an* Riemannian immersion *if $g^M = f^* g^N$. In words, if the metric on M induced by the immersion is equal to the existing metric on M. If additionally f is a diffeomorphism (bijective, smooth, smooth inverse) then we call f an* isometry*. Two Riemannian manifolds are* isometric *if there is a isometry between them.*

As above, if M is just a manifold and we have an immersion $f : M \to N$ to a Riemannian manifold, then we can endow M with the pullback metric. Then f becomes a Riemannian immersion by definition.

Example 3.9. Suppose that we have an Riemannian immersion $f : M \to \mathbb{R}^3$ and let $R : \mathbb{R}^3 \to \mathbb{R}^3$ be a rotation. Define $R \circ f : M \to \mathbb{R}^3$; this is also a Riemannian immersion, as we will now prove. The essential step of the calculation is to notice that $TR = R$ because R is a linear transformation, and that R a rotation doesn't change the inner product $g^{\mathbb{R}^3}$. Therefore

$$\begin{aligned}(R\circ f)^* g^{\mathbb{R}^3}(v,w) &= g^{\mathbb{R}^3}(T(R\circ f)v, T(R\circ f)w) = g^{\mathbb{R}^3}((TR\circ Tf)v, (TR\circ Tf)w)\\ &= g^{\mathbb{R}^3}(R(Tf(v)), R(Tf(w))) = g^{\mathbb{R}^3}(Tf(v), Tf(w))\\ &= f^* g^{\mathbb{R}^3}(v,w) = g^M(v,w).\end{aligned}$$

In the last line we used that f is a Riemannian immersion.

Exercise 3.10. Generalise the above example to prove: the composition of two Riemannian immersions is a Riemannian immersion.

A weaker condition to isometry is that of a conformal map.

Definition 3.11. *Let M, N be Riemannian manifolds and let $f : M \to N$ be an immersion. We say that f is* conformal *if there exists a smooth function $\lambda : M \to \mathbb{R}$ such that $g^M = \lambda f^* g^N$.*

A conformal map does not preserve lengths or distances, but it does preserve angles since

$$g^N(Tf(X), Tf(Y)) = f^* g^N(X,Y) = \lambda g^M(X,Y)$$

implies

$$\frac{g^N(Tf(X), Tf(Y))}{\|Tf(X)\|_{g^N}\|Tf(Y)\|_{g^N}} = \frac{\lambda g^M(X,Y)}{\sqrt{\lambda}\|X\|_{g^M}\sqrt{\lambda}\|Y\|_{g^M}} = \frac{g^M(X,Y)}{\|X\|_{g^M}\|Y\|_{g^M}}.$$

Example 3.12. Consider inverse stereographic projection $\Phi = \phi_N^{-1}$ as a function between the plane $\mathbb{R}^2$ with the standard metric and the sphere $\mathbb{S}^2$ with the induced metric of $\mathbb{R}^3$. Then Φ is not a Riemannian immersion, because the pullback metric $\Phi^* g^{\mathbb{S}^2}$ computed in Example 3.6 is not equal to the standard metric δ_{ij}. However, Φ is conformal because

$$\Phi^* g^{\mathbb{S}^2} = \frac{4}{(\|x\|^2 + 1)^2} \delta_{ij}.$$

Notice for example, that in stereographic coordinates the lines through the origin are lines of longitude and circles centered at the origin are lines of latitude, and these are always perpendicular to one another.

3.2 Quaternions and $\mathbb{S}^3$

In this section we introduce the quaternions as a means to understand the rotations of $\mathbb{S}^3$. The 3-sphere is a beautiful manifold because it is also a group. A manifold that is also a group is called a *Lie group*. We will not go into the general theory of Lie groups, but they come with a natural way to move vectors around, something we are trying to achieve in this chapter. The example of Lie groups is therefore very instructive for us.

The quaternions are a four dimensional real vector space $\{a_0+a_1\mathbf{i}+a_2\mathbf{j}+a_3\mathbf{k}\}$. A quaternion has a real part $\operatorname{Re} a = a_0$ and an imaginary part $\operatorname{Im} a = a_1\mathbf{i}+a_2\mathbf{j}+a_3\mathbf{k}$. Unlike for complex numbers, the imaginary part of a quaternion is not real. The quaternionic conjugate is $\bar{a} = \operatorname{Re} a - \operatorname{Im} a$. Clearly $\operatorname{Re}\bar{a} = \operatorname{Re} a$ and $\operatorname{Im}\bar{a} = -\operatorname{Im} a$. Elements of the subspace $\{a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}\}$ are called imaginary.

Famously the quaternions have an associative but non-commutative multiplication, defined by $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$ and 1 is the identity. We also use the notation $\mathbf{e} = 1$ to aid clarity. For example $\mathbf{ij} = \mathbf{k}$ because we multiply $\mathbf{ijk} = -1$ on the right by $\mathbf{k}$ to get $\mathbf{ijk}^2 = -\mathbf{k}$ and use $\mathbf{k}^2 = -1$. On the other hand $\mathbf{ji} = -\mathbf{k}$: from $\mathbf{ijk} = -1$ we get $1 = \mathbf{kji}$ and now multiply on the left by $\mathbf{k}$. This doesn't mean that every multiplication of quaternions is anti-commuting:

$$(1+\mathbf{i})(1+\mathbf{j}) = 1+1\mathbf{j}+\mathbf{i}1+\mathbf{ij} = 1+\mathbf{i}+\mathbf{j}+\mathbf{k},$$
$$(1+\mathbf{j})(1+\mathbf{i}) = 1+1\mathbf{i}+\mathbf{j}1+\mathbf{ji} = 1+\mathbf{i}+\mathbf{j}-\mathbf{k}.$$

According to legend on Monday 16 October 1843, as Hamilton was walking to the Royal Irish Academy, he had the idea that to define a multiplication on $\mathbb{R}^4$ it must be non-commutative, whereupon he carved the above equations into the side of Brougham Bridge. I have been to the bridge but was unable to find the carving, so instead I offer the following simple trick to remember the multiplication rule. Draw $\mathbf{i}, \mathbf{j}, \mathbf{k}$ on a directed circle. Multiplication of two elements gives the third, with a plus sign if they are in the correct direction and a minus sign if they are in the reverse direction. This is of course the same rule as for the cross product in $\mathbb{R}^3$.

A direct computation shows that $a\bar{a} = \bar{a}a = a_0^2 + a_1^2 + a_2^2 + a_3^2$ is always real and non-negative. Thus we can define the norm $|a| = \sqrt{a\bar{a}}$. The norm shows that every non-zero quaternion has a two-sided inverse, namely $a^{-1} = |a|^{-2}\bar{a}$. Therefore the quaternions are a non-commutative field.

Exercise 3.13. Prove the following:

a. The dot product can be calculated as $a \cdot b = \operatorname{Re}(\bar{a}b)$,

b. $a\bar{a} = \bar{a}a = a_0^2 + a_1^2 + a_2^2 + a_3^2$,

c. Conjugation is order reversing $\overline{ab} = \bar{b}\,\bar{a}$,

d. The norm is multiplicative $|ab| = |a|\,|b|$.

This norm is plainly the same as the usual norm on $\mathbb{R}^4$. The unit quaternions (those with norm 1) are as a set $\mathbb{S}^3 \subset \mathbb{R}^4$. Therefore the 3-sphere is a Lie group, because we can multiply two elements of it together in a way that can be undone. This is rather special, the only spheres that are Lie groups are $\mathbb{S}^0$ ($\mathbb{S}^0 = \{\pm 1\}$ in $\mathbb{R}^1$), $\mathbb{S}^1$ (add the angles), and $\mathbb{S}^3$.

If we choose $a \in \mathbb{S}^3$ we can look at the function $L_a : \mathbb{S}^3 \to \mathbb{S}^3$ defined by $L_a(q) = aq$ This is a bijective function, because the inverse is $L_{a^{-1}}$. And $L_a(\mathbf{e}) = a\mathbf{e} = a$. Therefore the tangent map of L_a takes $T_{\mathbf{e}}\mathbb{S}^3$ to $T_a\mathbb{S}^3$. Moreover, the tangent map is also bijective: from the chain rule

$$\mathrm{id}_{T_{\mathbf{e}}\mathbb{S}^3} = T(L_a \circ L_{a^{-1}}) = TL_a \circ TL_{a^{-1}}$$

Indeed, this inverse has the property that it takes a to the identity $L_{a^{-1}}(a) = a^{-1}a = \mathbf{e}$. This gives us a way to move any tangent vector of $\mathbb{S}^3$ to $T_{\mathbf{e}}\mathbb{S}^3$. Just as in Example 2.26, this shows us that $T\mathbb{S}^3$ is trivial. The function $T_aL_{a^{-1}} : T_a\mathbb{S}^3 \to T_{\mathbf{e}}\mathbb{S}^3$ is called the *left trivialisation*. Likewise we can define $R_a(q) = qa$ and we have the *right trivialisation* $T_aR_{a^{-1}} : T_a\mathbb{S}^3 \to T_{\mathbf{e}}\mathbb{S}^3$

Example 3.14. Let us compute these for the point $a = \mathbf{i} = (0, 1, 0, 0) \in \mathbb{S}^3$. The inverse of a is $a^{-1} = -\mathbf{i}$, since $\mathbf{i}(-\mathbf{i}) = 1$. If we have any point $q = q^0 + q^1\mathbf{i} + q^2\mathbf{j} + q^3\mathbf{k}$ then

$$L_{a^{-1}}(q) = (-\mathbf{i})(q^0 + q^1\mathbf{i} + q^2\mathbf{j} + q^3\mathbf{k}) = q^1 - q^0\mathbf{i} + q^3\mathbf{j} - q^2\mathbf{k}.$$

This does indeed have the property that $L_{a^{-1}}(a) = 1 - 0 + 0 - 0 = 1 = \mathbf{e}$. Next we use some geometry to avoid using charts. We know that the tangent vectors in $T_a\mathbb{S}^3$ are perpendicular to a, because this is a sphere. We write

$$T_a\mathbb{S}^3 = \{v^1\mathbf{e} + v^2\mathbf{j} + v^3\mathbf{k} \mid v^1, v^2, v^3 \in \mathbb{R}\}.$$

Because $L_{a^{-1}}(q)$ is linear in q, we know

$$T_aL_{a^{-1}}(v^1\mathbf{e} + v^2\mathbf{j} + v^3\mathbf{k}) = -v^1\mathbf{i} + v^3\mathbf{j} - v^2\mathbf{k}.$$

For the right trivialisation

$$R_{a^{-1}}(q) = (q^0 + q^1\mathbf{i} + q^2\mathbf{j} + q^3\mathbf{k})(-\mathbf{i}) = q^1 - q^0\mathbf{i} - q^3\mathbf{j} + q^2\mathbf{k},$$
$$T_aR_{a^{-1}}(v^1\mathbf{e} + v^2\mathbf{j} + v^3\mathbf{k}) = -v^1\mathbf{i} - v^3\mathbf{j} + v^2\mathbf{k}.$$

So these two trivialisations on $\mathbb{S}^3$ are different from one another.

Example 3.15. We can generalise the previous example to work for any point $a \in \mathbb{S}^3$. Just like i is a right-angle rotation of the complex plane, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are all right-angle rotations of the quaternions. Therefore $a\mathbf{i}$, $a\mathbf{j}$, $a\mathbf{k}$ is an orthonormal basis of $T_a\mathbb{S}^3$. Alternatively, since

$$L_a(q) = a(q^0 + q^1\mathbf{i} + q^2\mathbf{j} + q^3\mathbf{k}) = q^0a + q^1a\mathbf{i} + q^2a\mathbf{j} + q^3a\mathbf{k}$$

and $\mathbf{i}, \mathbf{j}, \mathbf{k}$ is a basis for $T_{\mathbf{e}}\mathbb{S}^3$ we know that

$$T_{\mathbf{e}}L_a(v^1\mathbf{i} + v^2\mathbf{j} + v^3\mathbf{k}) = v^1a\mathbf{i} + v^2a\mathbf{j} + v^3a\mathbf{k}$$

is all of $T_a\mathbb{S}^3$. This shows us that identifying $T_a\mathbb{S}^3$ with $T_qe\mathbb{S}^3$ is the same as writing it with respect to the pushforward of a basis. If $v \in T_a\mathbb{S}^3$ then we get

$$T_aL_{a^{-1}}v = a^{-1}v.$$

We call the vector field on $\mathbb{S}^3$ a *left-invariant* field when it has the form

$$X|_a = av$$

for $v \in T_{\mathbf{e}}\mathbb{S}^3$, because every vector $X|_a$ corresponds to v using the left trivialisation. Ditto we have the *right-invariant* vector fields

$$Y|_a = va.$$

Example 3.16.

$$X|_a = a\mathbf{i} = -a_1\mathbf{e} + a_0\mathbf{i} + a_3\mathbf{j} - a_2\mathbf{k}$$

is a left-invariant vector field on $\mathbb{S}^3$. We recognise $X|_{\mathbf{i}} = -\mathbf{e}$ as a vector in $T_{\mathbf{i}}\mathbb{S}^3$.

3.3 Covariant Derivatives

We have seen numerous examples thus far of how we cannot simply move vectors around in a chart like we can in euclidean space. If you take a tangent vector at one point of the sphere and translate it in $\mathbb{R}^3$ to another point of the sphere, it may not be tangent anymore. As we observed following Example 2.18, a vector might have the same coordinates at different points in one chart, but not in another. And in Example 3.3 we saw that one coordinate basis vector changed its length as you moved around, while the other stayed the same length.

There is also a common thought experiment. Suppose that you are standing on the equator facing east. You walk forward without turning, until you have walked half way around the Earth. Then, still without turning, you begin to sidestep to the north. You sidestep all the way to the north pole, but keep going until you have returned to your original position. The remarkable fact is, even though at no stage did you turn, you are now facing west.

Exercise 3.17. Can you modify the journey so that you end up facing other directions? What is the connection between the area your journey encompasses and the final rotation angle?

However, the naive definition of the derivative of a vector field

$$\lim_{h\to 0} \frac{1}{h}(X|_{p+h} - X|_p)$$

asks us to subtract two vectors at different points. Indeed, any non-trivial definition of a derivative of a vector field is going to require us to compare vectors at different points. Geometrically, thinking about a surface, what we want to do is to 'roll' the tangent plane along the surface to another point. This idea is called development and the relation between two tangent planes was called an *affine connection*, because it was an affine transformation of one plane to another. In modern terminology it is more common to call this a *parallel transport operator*, for reasons that will be explained in Section 3.4. Already from the above thought experiment we see that a parallel transport operator will depend not just on the two start and end points, but on the path between those points.

The modern approach, which we will ultimately take, uses a different point of view. It asks: how much are vector fields are changing? Once we have a basis of vector fields and we know their changes, then we can measure all other vector fields against them. This leads to the definition of a *covariant derivative*, a type of differential operator on vector fields. It is extremely common to call this an *connection*, but we will refrain from doing so, at least until we have made clear the relationship with the parallel transport operator. Though the two approaches are equivalent, the modern approach is the much easier place to begin. On the other hand, some of the definitions and motivations for the modern approach only really make sense from the point of view of the traditional approach.

Definition 3.18. *A* covariant derivative *on a manifold M is a function ∇ that acts on two vector fields to produce a third. We write it as $\nabla_X Y$, with X being the 'direction'. It has the following properties for all smooth functions $f : M \to \mathbb{R}$ and vector fields $X, \tilde{X}, Y, \tilde{Y}$:*

a. It is C^∞-linear in the direction:

$$\nabla_{fX+\tilde{X}} Y = f\nabla_X Y + \nabla_{\tilde{X}} Y.$$

b. It is additive in the derivative:

$$\nabla_X(Y+\tilde{Y}) = \nabla_X Y + \nabla_X \tilde{Y}.$$

c. It obeys the product (Leibniz) rule:

$$\nabla_X(fY) = X(f)Y + f\nabla_X Y.$$

Example 3.19. Consider euclidean space $\mathbb{R}^n$ and let $X = X^i\partial_i, Y = Y^i\partial_i$ be vector fields in the chart x^i. Then

$$\nabla^{\text{euc}}_X Y := X^i \frac{\partial Y^j}{\partial x^i}\partial_j$$

is a covariant derivative.

You might be confused, because in Example 2.31 we said this formula didn't work. Indeed, this formula is not chart independent. This definition is saying explicitly "use this particular coordinates to do the derivative and not others". If you write this covariant derivative in polar coordinates, then the formula for this covariant derivative will look different. But this is why we say that it is a covariant derivative, we are not claiming uniqueness.

Exercise 3.20. Check the above example has the three properties that are required of a covariant derivative.

The above example suggests that there are many covariant derivatives on a manifold. At least for a manifold that can be covered by a single chart, every set of coordinates gives a covariant derivative. In the following theorem we characterise the set of covariant derivatives.

Theorem 3.21 (Tensorial)**.** *Let ∇^0, ∇^1 be two covariant derivatives derivatives. Define their difference $A(X,Y) := \nabla^0_X Y - \nabla^1_X Y$. Then A is C^∞-linear in both X and Y.*

Proof. C^∞-linear in X is immediate from Property a of covariant derivatives. C^∞-linear in Y is not too much harder to show, we use Properties b and c:

$$\begin{aligned} A(X, fY+\tilde{Y}) &= \nabla^0_X(fY) + \nabla^0_X\tilde{Y} - \nabla^1_X(fY) - \nabla^1_X\tilde{Y} \\ &= X(f)Y + f\nabla^0_X Y - X(f)Y - \nabla^1_X Y + A(X,\tilde{Y}) \\ &= fA(X,Y) + A(X,\tilde{Y}). \end{aligned}$$

□

Exercise 3.22. Prove the converse of Theorem 3.21: Let ∇ is a covariant derivative on M. For all vector fields X, Y let $A(X,Y)$ be a smooth vector field. Suppose that this function A is C^∞-linear in both X, Y. Then $\tilde{\nabla} := \nabla + A$ is also a covariant derivative.

Corollary 3.23 (Affineness)**.** *The space of covariant derivatives on M is affine in the following sense: if $t \in \mathbb{R}$ is a constant and ∇^0, ∇^1 are two covariant derivatives, so is $\nabla^t := (1-t)\nabla^0 + t\nabla^1$.*

Proof. Observe that $\nabla^t = \nabla^0 + t(\nabla^1 - \nabla^0)$. The corollary now follows from Theorem 3.21 and its converse Exercise 3.22. □

The above theorems give us a way to construct new covariant derivatives from existing ones (and in fact construct every covariant derivative). But we need one to start with. One can prove[1] that every manifold has a covariant derivative, but the proof is technical and not practically useful. We have seen in Example 3.19 that if one chart covers the whole space, then we can declare is special and use the directional derivative. For manifolds that are a submanifold of a bigger space, the following example is typical.

Example 3.24. Consider the sphere $\mathbb{S}^2$ inside $\mathbb{R}^3$. We can understand any vector field Y on $\mathbb{S}^2$ as a function $X : \mathbb{S}^2 \to \mathbb{R}^3$ using the pushforward. Therefore we can differentiate X as an $\mathbb{R}^3$ valued function in the usual way. The trouble is that the directional derivative $X^i \frac{\partial Y^j}{\partial p^i} \frac{\partial}{\partial p^j}$ might no longer be tangent to $\mathbb{S}^2$. Therefore this does not meet the definition of a covariant derivative on $\mathbb{S}^2$. What we can do however is to project this directional derivative onto the tangent space. We define the *tangent covariant derivative* as

$$\nabla^\top_X Y = \operatorname*{proj}_{T_p\mathbb{S}^2} X^i \frac{\partial Y^j}{\partial p^i} \frac{\partial}{\partial p^j}.$$

Let's check the three required properties. The two linearity properties just follow from the linearity of the projection

$$\begin{aligned}
\nabla^\top_{fX+\tilde{X}} Y &= \operatorname*{proj}_{T_p\mathbb{S}^2} \left(f X^i \frac{\partial Y^j}{\partial p^i} \frac{\partial}{\partial p^j} + \tilde{X}^i \frac{\partial Y^j}{\partial p^i} \frac{\partial}{\partial p^j} \right) \\
&= f \operatorname*{proj}_{T_p\mathbb{S}^2} X^i \frac{\partial Y^j}{\partial p^i} \frac{\partial}{\partial p^j} + \operatorname*{proj}_{T_p\mathbb{S}^2} \tilde{X}^i \frac{\partial Y^j}{\partial p^i} \frac{\partial}{\partial p^j} = f\nabla^\top_X Y + \nabla^\top_{\tilde{X}} Y, \\
\nabla^\top_X (Y + \tilde{Y}) &= \operatorname*{proj}_{T_p\mathbb{S}^2} \left(X^i \frac{\partial Y^j}{\partial p^i} \frac{\partial}{\partial p^j} + X^i \frac{\partial \tilde{Y}^j}{\partial p^i} \frac{\partial}{\partial p^j} \right) = \nabla^\top_X Y + \nabla^\top_X \tilde{Y}.
\end{aligned}$$

For the third property, we need to recognise that $X(f)Y$ is already tangent to $\mathbb{S}^2$, so the projection leaves it unaltered:

$$\begin{aligned}
\nabla^\top_X (fY) &= \operatorname*{proj}_{T_p\mathbb{S}^2} X^i \frac{\partial (fY^j)}{\partial p^i} \frac{\partial}{\partial p^j} = \operatorname*{proj}_{T_p\mathbb{S}^2} \left(X(f)Y + X^i f \frac{\partial Y^j}{\partial p^i} \frac{\partial}{\partial p^j} \right) \\
&= X(f)Y + f\nabla^\top_X Y.
\end{aligned}$$

Next we examine what type of derivative a covariant derivative is. We will show that it is a directional derivative, in a sense that will be developed. To this end, the first property to notice is that although the direction and the derived vector fields have dramatically different behaviour under scaling by a smooth function, they are both $\mathbb{R}$-linear. If a is a constant then

$$\nabla_{aX} Y = a\nabla_X Y, \qquad \nabla_X (aY) = X(a)Y + a\nabla_X Y = a\nabla_X Y.$$

Consequently, if either field is zero, then so is the covariant derivative. Moreover, using cut-off functions, the covariant derivative only depends on local information.[2] In fact something stronger is true of X:

[1] Lee Proposition 4.5

[2] See Lee Lemma 4.1 for a proof. We prove a stronger statement in Lemma 3.26.

Lemma 3.25 (Directional Derivative)**.** *The value of $\nabla_X Y$ at $p \in M$ only depends on $X|_p$ and not other values of X.*

Proof. By linearity, it suffices to prove that $X|_p = 0$ implies $(\nabla_X Y)|_p = 0$. Writing X in a chart we have $X = X^i \partial_i$ and $X^i(p) = 0$ for all the coefficients. Then

$$\big(\nabla_{X^i\partial_i} Y\big)|_p = \big(X^i \nabla_{\partial_i} Y\big)\Big|_p = X^i(p)\big(\nabla_{\partial_i} Y\big)\Big|_p = 0. \qquad \square$$

For this reason we sometimes speak of the covariant derivative $\nabla_v Y$ in a direction $v \in T_pM$. The same is not true for Y: the covariant derivative really is a derivative of Y and depends on its values in a neighbourhood of a point. However, to compute $\nabla_v Y$ you don't need to know Y completely on an open neighbourhood of p, it is enough to know Y on a curve whose tangent is v.

Lemma 3.26 (Curve Derivative)**.** *Let $Y, \tilde{Y}$ be two vector fields and let $\alpha : (a, b) \to M$ be a smooth curve with $\alpha(0) = p$ and $\alpha'(0) = v$. Suppose that $Y \circ \alpha = \tilde{Y} \circ \alpha$. Then $\nabla_v Y = \nabla_v \tilde{Y}$.*

Proof. Let us consider the situation in a chart, writing $v = v^i \partial_i|_p$, $Y = Y^i \partial_i$ and $\tilde{Y} = \tilde{Y}^i \partial_i$. Then by the properties of covariant derivatives,

$$\nabla_v Y = \nabla_{v^i \partial_i|_p}(Y^j \partial_j) = v^i \nabla_{\partial_i|_p}(Y^j \partial_j) = v^i \left.\frac{\partial Y^j}{\partial x^i}\right|_p \partial_j + v^i Y^j(p) \nabla_{\partial_i|_p} \partial_j,$$

and likewise for $\tilde{Y}$. Now, Y and $\tilde{Y}$ agree on α, so $Y(p) = \tilde{Y}(p)$. Moreover, by the chain rule

$$v^i \left.\frac{\partial Y^j}{\partial x^i}\right|_p = \left.\frac{d}{dt}(Y^j \circ \alpha)\right|_p = \left.\frac{d}{dt}(\tilde{Y}^j \circ \alpha)\right|_p = v^i \left.\frac{\partial \tilde{Y}^j}{\partial x^i}\right|_p.$$

Hence

$$\nabla_v Y = v^i \left.\frac{\partial Y^j}{\partial x^i}\right|_p \partial_j + v^i Y^j(p) \nabla_{\partial_i|_p} \partial_j = v^i \left.\frac{\partial \tilde{Y}^j}{\partial x^i}\right|_p \partial_j + v^i \tilde{Y}^j(p) \nabla_{\partial_i|_p} \partial_j = \nabla_v \tilde{Y} \qquad \square$$

This lemma tells us that we can really view the covariant derivative as a generalisation of a directional derivative. This is in contrast to other derivatives of vector fields. Recall Example 2.34. Now consider the vector fields from that example along the curve $\alpha(t) = (t, 0)$, the x-axis. We have $X \circ \alpha = \partial_1$, $Y \circ \alpha = \partial_2$, and $V \circ \alpha = \partial_1$. But $[X, Y] = 0$ while $[V, Y] = -\partial_1$. This shows that the Lie bracket is not a covariant derivative.

To break up all this theory, let's do another example.

Example 3.27. We define a covariant derivative ∇^L on $\mathbb{S}^3$ in the following way. Given any vector field Y on $\mathbb{S}^3$, use left trivialisation to write it as a function $\tilde{Y} : \mathbb{S}^3 \to T_{\mathbf{e}}\mathbb{S}^3$. From Example 3.15 we know this has the formula $p \mapsto p^{-1} Y|_p$ using quaternions. Now that we have a function to the same vector space, there is no problem differentiating. This gives us a function $X(\tilde{Y}) : M \to T_{\mathbf{e}}\mathbb{S}^3$. Use the left trivialisation again to move the result back to $T_p\mathbb{S}^3$.

Putting this all in one formula gives

$$(\nabla^L_X Y)|_p := (T_{\mathbf{e}} L_p \circ X \circ T_p L_{p^{-1}}) Y.$$

This covariant derivative has the property that the derivative of a left-invariant vector field is always zero. This is because, by definition, after you bring its vectors to $\mathbf{e}$ they are all the same. In other words $\tilde{Y}$ is constant and thus has zero derivative.

So to see an interesting example, we need to use a non-left-invariant vector field. Consider $Y|_p = \mathbf{i}p$. We know that $\tilde{Y}(p) = p^{-1}\mathbf{i}p$. To proceed we need to choose a direction field X. We know that the value of the covariant derivative at any point only depends on the value of X at that point. So for simplicity let us calculate for the point $\mathbf{i}$ in the direction $\mathbf{j} = \frac{\partial}{\partial p^2}$:

$$\begin{aligned} X|_{\mathbf{i}}\tilde{Y} &= \frac{\partial}{\partial p^2} p^{-1}\mathbf{i}p\Big|_{\mathbf{i}} = -p^{-1}\frac{\partial p}{\partial p^2}p^{-1}\mathbf{i}p + p^{-1}\mathbf{i}\frac{\partial p}{\partial p^2}\Big|_{\mathbf{i}} = -p^{-1}\mathbf{j}p^{-1}\mathbf{i}p + p^{-1}\mathbf{i}\mathbf{j}\big|_{\mathbf{i}} \\ &= -\mathbf{i}^{-1}\mathbf{j}\mathbf{i}^{-1}\mathbf{i}\mathbf{i} + \mathbf{i}^{-1}\mathbf{i}\mathbf{j} = \mathbf{j} + \mathbf{j} = 2\mathbf{j}. \end{aligned}$$

Finally, we move this back to $T_{\mathbf{j}}\mathbb{S}^3$

$$(\nabla_X Y)|_{\mathbf{j}} = T_{\mathbf{e}}L_{\mathbf{j}}(2\mathbf{j}) = \mathbf{j}\,2\mathbf{j} = -2\mathbf{e}.$$

In the same manner, we can define a covariant derivative ∇^R using the right trivialisation.

In the examples above, to define a covariant derivative we really gave a directional derivative. But what is the minimal information required to specify a covariant derivative? Because covariant derivatives are local, we give the answer in a chart. Let ∂_i be the coordinate vector fields. Then for each pair i, j we have a vector field $\nabla_{\partial_i}\partial_j$. This vector field must be able to be written

$$\nabla_{\partial_i}\partial_j = \Gamma^k_{ij}\partial_k,$$

for some coefficients Γ^k_{ij}. These coefficients are called *Christoffel coefficients*, though be aware that some authors reserve this name for a special case. This is sufficient information to determine ∇ because

$$\nabla_X Y = X^i\nabla_{\partial_i}(Y^j\partial_j) = X^i\frac{\partial Y^j}{\partial x^i}\partial_j + X^iY^j\nabla_{\partial_i}\partial_j = \left(X^i\frac{\partial Y^k}{\partial x^i} + X^iY^j\Gamma^k_{ij}\right)\partial_k.$$

Example 3.28. Let us consider $\mathbb{R}^2$ with ∇^{euc}. We see by comparison of its definition in Example 3.19 with the formula above that Γ^k_{ij} is zero for all points and all indices in the standard chart.

But let us compute it with respect to polar coordinates. By the definition of ∇^{euc}, we have to calculate in the x^1, x^2 coordinates. We have

$$\begin{aligned} \frac{\partial}{\partial r} &= \cos\theta\frac{\partial}{\partial x^1} + \sin\theta\frac{\partial}{\partial x^2} = \frac{x^1}{\sqrt{(x^1)^2+(x^2)^2}}\frac{\partial}{\partial x^1} + \frac{x^2}{\sqrt{(x^1)^2+(x^2)^2}}\frac{\partial}{\partial x^2} \\ \frac{\partial}{\partial\theta} &= -r\sin\theta\frac{\partial}{\partial x^1} + r\cos\theta\frac{\partial}{\partial x^2} = -x^2\frac{\partial}{\partial x^1} + x^1\frac{\partial}{\partial x^2}. \end{aligned}$$

Hence we can calculate

$$\begin{aligned}\nabla_{\frac{\partial}{\partial r}}\frac{\partial}{\partial\theta} &= \frac{x^1}{\sqrt{(x^1)^2+(x^2)^2}}\nabla_{\frac{\partial}{\partial x^1}}\frac{\partial}{\partial\theta} + \frac{x^2}{\sqrt{(x^1)^2+(x^2)^2}}\nabla_{\frac{\partial}{\partial x^2}}\frac{\partial}{\partial\theta}\\ &= \frac{x^1}{\sqrt{(x^1)^2+(x^2)^2}}\left(\frac{\partial(-x^2)}{\partial x^1}\frac{\partial}{\partial x^1}+\frac{\partial x^1}{\partial x^1}\frac{\partial}{\partial x^2}\right)\\ &\quad+\frac{x^2}{\sqrt{(x^1)^2+(x^2)^2}}\left(\frac{\partial(-x^2)}{\partial x^2}\frac{\partial}{\partial x^1}+\frac{\partial x^1}{\partial x^2}\frac{\partial}{\partial x^2}\right)\\ &= \frac{x^1}{\sqrt{(x^1)^2+(x^2)^2}}\frac{\partial}{\partial x^2}-\frac{x^2}{\sqrt{(x^1)^2+(x^2)^2}}\frac{\partial}{\partial x^1}\\ &= -\sin\theta\frac{\partial}{\partial x^1}+\cos\theta\frac{\partial}{\partial x^2}=\frac{1}{r}\frac{\partial}{\partial\theta},\end{aligned}$$

and hence in polar coordinates

$$\Gamma^r_{r,\theta}=0,\qquad \Gamma^\theta_{r,\theta}=\frac{1}{r}.$$

The other six coefficients are calculated similarly.

Example 3.29. Let's calculate the Christoffel coefficients for $\mathbb{S}^2$ with the connection $\nabla^\top$ from Example 3.24 in the chart U_N. Because the definition of $\nabla^\top$ uses the geometry of $\mathbb{R}^3$ we need the pushforwards of the coordinate basis vectors, but we computed those in Example 2.21 and Example 3.6. Denote $E_i = J_x(\phi_N^{-1})\frac{\partial}{\partial x^i}$ for convenience. Before we dive into calculations, we can simplify things a little by recognising

$$X^i\frac{\partial Y^j}{\partial p^i}=X(Y^j).$$

This is useful because we can apply the directional derivative version of the pushforward

$$E_i(Y^j)=T\phi_N^{-1}\left(\frac{\partial}{\partial x^i}\right)(Y^j)=\frac{\partial}{\partial x^i}(Y^j\circ\phi_N^{-1}).$$

Therefore the covariant derivative is

$$\nabla^\top_{\partial_i}\partial_j=\operatorname*{proj}_{T_p\mathbb{S}^3}\frac{\partial}{\partial x^i}(E_j^k\circ\phi_N^{-1})\frac{\partial}{\partial p^k}$$

We gather some intermediate steps first. Guided by longitude and latitude, we have

$$E_1=\frac{2}{(\|x\|^2+1)^2}\begin{pmatrix}-(x^1)^2+(x^2)^2+1\\-2x^1x^2\\2x^1\end{pmatrix}\quad x^1E_1+x^2E_2=\frac{2}{(\|x\|^2+1)^2}\begin{pmatrix}-x^1(\|x\|^2-1)\\-x^2(\|x\|^2-1)\\2\|x\|^2\end{pmatrix}$$

$$E_2=\frac{2}{(\|x\|^2+1)^2}\begin{pmatrix}-2x^1x^2\\(x^1)^2-(x^2)^2+1\\2x^2\end{pmatrix}\quad x^2E_1-x^1E_2=\frac{2}{\|x\|^2+1}\begin{pmatrix}x^2\\-x^1\\0\end{pmatrix}.$$

With an eye on what's to come:

$$p = \phi_N^{-1} = \frac{1}{\|x\|^2+1}\begin{pmatrix} 2x^1 \\ 2x^2 \\ (x^1)^2+(x^2)^2-1 \end{pmatrix}$$
$$-p + \frac{1}{2}(\|x\|^2+1)(x^1E_1 + x^2E_2) = \begin{pmatrix} -x^1 \\ -x^2 \\ 1 \end{pmatrix}.$$

So then the derivatives are

$$\begin{aligned}\frac{\partial}{\partial x^1}E_1 &= \frac{-4(2x^1)}{(\|x\|^2+1)^3}\begin{pmatrix} -(x^1)^2+(x^2)^2+1 \\ -2x^1x^2 \\ 2x^1 \end{pmatrix} + \frac{4}{(\|x\|^2+1)^2}\begin{pmatrix} -x^1 \\ -x^2 \\ 1 \end{pmatrix} \\ &= \frac{-4x^1}{\|x\|^2+1}E_1 + \frac{4}{(\|x\|^2+1)^2}\left[-p + \frac{1}{2}(\|x\|^2+1)(x^1E_1+x^2E_2)\right] \\ \frac{\partial}{\partial x^2}E_1 &= \frac{-4(2x^2)}{(\|x\|^2+1)^3}\begin{pmatrix} -(x^1)^2+(x^2)^2+1 \\ -2x^1x^2 \\ 2x^1 \end{pmatrix} + \frac{4}{(\|x\|^2+1)^2}\begin{pmatrix} x^2 \\ -x^1 \\ 0 \end{pmatrix} \\ &= \frac{-4x^2}{\|x\|^2+1}E_1 + \frac{2}{\|x\|^2+1}\left[x^2E_1 - x^1E_2\right].\end{aligned}$$

Similarly

$$\begin{aligned}\frac{\partial}{\partial x^1}E_2 &= \frac{-4x^1}{\|x\|^2+1}E_2 - \frac{2}{\|x\|^2+1}\left[x^2E_1 - x^1E_2\right] \\ \frac{\partial}{\partial x^2}E_2 &= \frac{-4x^2}{\|x\|^2+1}E_2 + \frac{4}{(\|x\|^2+1)^2}\left[-p + \frac{1}{2}(\|x\|^2+1)(x^1E_1+x^2E_2)\right].\end{aligned}$$

Because we have be able to write these derivative as a combination of vectors in $T_p\mathbb{S}^2$ and p, which is perpendicular to $T_p\mathbb{S}^2$, we can take the projection by killing the p part. For example

$$\begin{aligned}\nabla^\top_{\partial_1}\partial_1 &= \frac{-4x^1}{\|x\|^2+1}E_1 + \frac{4}{(\|x\|^2+1)^2}\left[-0 + \frac{1}{2}(\|x\|^2+1)(x^1E_1+x^2E_2)\right] \\ &= \frac{-2x^1}{\|x\|^2+1}E_1 + \frac{2x^2}{\|x\|^2+1}E_2\end{aligned}$$

From this we read that

$$\Gamma^1_{11} = \frac{-2x^1}{\|x\|^2+1}, \qquad \Gamma^2_{11} = \frac{2x^2}{\|x\|^2+1}.$$

For the other derivative of E_1, the projection is trivial, and

$$\Gamma^1_{21} = \frac{-2x^2}{\|x\|^2+1}, \qquad \Gamma^2_{21} = \frac{-2x^1}{\|x\|^2+1}.$$

And from the derivatives of E_2 we obtain:

$$\Gamma^1_{12} = \frac{-2x^2}{\|x\|^2+1}, \qquad \Gamma^2_{12} = \frac{-2x^1}{\|x\|^2+1},$$
$$\Gamma^1_{22} = \frac{2x^1}{\|x\|^2+1}, \qquad \Gamma^2_{22} = \frac{-2x^2}{\|x\|^2+1}.$$

Exercise 3.30 (Lee Lemma 4.4)**.** Suppose that M is a manifold covered by a single chart U. Show that the set of covariant derivatives on M is in one-to-one correspondence with the set of Christoffel coefficients. That is, show that every choice of n^3 functions Γ^k_{ij} gives a covariant derivative.

Exercise 3.31. Derive the transformation formula for Γ^k_{ij} between two charts. Observe that it is neither covariant nor contravariant.

3.4 Parallel Transport

We began Section 3.3 with the motivation that we want to compare different tangent spaces to one another and a thought experiment about walking around the Earth. Then we went on to define covariant derivatives. Now it is time to connect the two (pardon the pun).

Definition 3.32. *Let M be a manifold with a connection ∇, $\alpha : (a, b) \to M$ a smooth curve and Y a vector field. We say that Y is* parallel *along α (with respect to ∇) if $\nabla_{\alpha'} Y = 0$ at all points on the curve.*

The inspiration of the name parallel is that the vectors of the vector field at different points are meant to be (in some sense) parallel to one another. Phrased different: we have a field of parallel vectors. Even though α' is not a vector field on M, this is well-defined due to Lemma 3.25. Similarly, we really only need to values of Y along the curve α to compute this condition, due to Lemma 3.26. Therefore any books build a theory of 'vector fields on curves'. We will avoid this extra theory by assuming the main result: so long as the curve α is injective and not pathological, every vector field on α can be extended to a vector field on M.

In a chart we have $\alpha'(t) = \frac{d\alpha^i}{dt}\partial_i$, so the condition becomes

$$0 = \left(\frac{d\alpha^i}{dt} \frac{\partial Y^k}{\partial x^i} + \Gamma^k_{ij} \frac{d\alpha^i}{dt} Y^j \right) \partial_j = \left(\frac{dY^k}{dt} + \Gamma^k_{ij} \frac{d\alpha^i}{dt} Y^j \right) \partial_j, \tag{3.33}$$

where we treat the vector field as a function of t, i.e. $Y(\alpha(t))$. Since Γ^k_{ij} and $\frac{d\alpha^i}{dt}$ are specified, we treat this as a system of ODEs for the functions $Y^i(t) : (a, b) \to \mathbb{R}$. By the uniqueness of solutions to ODEs, a parallel vector field is uniquely determined by its value at one point of the curve. Conversely by the existence of solutions to ODEs, given a vector $v \in T_{\alpha(t_0)}$ there exists a unique parallel field Y along α with $Y(t_0) = v$.

Let us revisit Example 3.24 and bring in our thought experiment. We can expand the thought experiment in the following way: while we are walking around the world without turning, we are holding a stick. The stick represents a vector field along the curve of our journey. Suppose at the start of our journey, the stick is pointing south (recall we are facing east). As we walk east around the world, our stick will continue to point south. Thus we ask whether the vector field $Y|_p = (0, 0, -1) \in T_p\mathbb{S}^2$ is parallel with respect to $\nabla^\top$ along the equator $\alpha(t) = (\cos t, \sin t, 0)$. Indeed it is, since Y is constant with respect to p,

$$\nabla^\top_{\alpha'} Y = \operatorname*{proj}_{T_p\mathbb{S}^2} \left(-\sin t \, \frac{\partial Y^j}{\partial p^1} \frac{\partial}{\partial p^j} + \cos t \, \frac{\partial Y^j}{\partial p^2} \frac{\partial}{\partial p^j} + 0 \, \frac{\partial Y^j}{\partial p^3} \frac{\partial}{\partial p^j} \right) = 0.$$

Now what about the original thought experiment? This time as we walk around the world, let the stick point forward. Clearly, if we don't turn, it should continue to point forward. In other words

$$Y = \alpha'(t) = (-\sin t, \cos t, 0). = (-p^2, p^1, 0).$$

This is not constant as a function into $\mathbb{R}^3$. Now when we compute

$$\nabla^\top_{\alpha'} Y = \operatorname*{proj}_{T_p\mathbb{S}^2} \left(-\sin t \, (1) \frac{\partial}{\partial p^2} + \cos t \, (-1) \frac{\partial}{\partial p^1} + 0 \right) = \operatorname*{proj}_{T_p\mathbb{S}^2} (-p) = 0.$$

We see from the calculation that the derivative of Y along the curve points towards the center of the sphere, so when projected to the tangent plane it becomes zero. In summary, parallel transport by $\nabla^\top$ on the sphere matches our intuition of 'walking without turning'. Of course there are many other covariant derivatives on the sphere, and with respect to them perhaps these two vector fields are not parallel.

Example 3.34. Let us consider the covariant derivative ∇^L on $\mathbb{S}^3$ from Example 3.27. We noted there that left-invariant vector fields have ∇^L-derivative zero at any point and in any direction. Hence left-invariant fields are parallel along every curve in $\mathbb{S}^3$.

Conversely, suppose Y is parallel along α. It follows from the definition of ∇^L that $t \mapsto T_{\alpha(t)}L_{\alpha(t)^{-1}}Y(\alpha(t))$ is constant. In words, if we consider Y as a function of t, ie $Y(\alpha(t))$ and move the vectors to $\mathbf{e}$ using the tangent map of the left action, ie $T_{\alpha(t)}L_{\alpha(t)^{-1}}$, then this function is constant. Though we don't have a formal definition, it is fair to say that Y is left-invariant along the curve.

The final observation for this example is that given any vector $w \in T_p\mathbb{S}^3$ there is a unique left-invariant vector field Y with $Y|_p = w$. Let $v = T_pLp^{-1}w$. Then $Y|_p = pv$ is the field. Therefore there is a unique way to parallel transport any vector to any other point of $\mathbb{S}^3$. Manifolds with this property are called *parallelisable*. It is equivalent to having a trivial tangent bundle.

In the above example, we encountered the idea of taking a vector v at one point $\alpha(t_0)$, finding a vector field Y with $Y|_{\alpha(t_0)}$ that is parallel along α, and in particular calculating the parallel vector at another point $w = Y|_{\alpha(t_1)}$. We call w the *parallel transport* of v along α. This is a function $P(\alpha)_t^s : T_{\alpha(t)}M \to T_{\alpha(s)}M$ called the *parallel transport operator*. Because the ODE is linear in Y, the parallel transport operator is linear: If Y is the parallel vector field with $Y|_{\alpha(t)} = v$ and $\tilde{Y}$ is the parallel vector field with $\tilde{Y}|_{\alpha(t)} = \tilde{v}$, then $Y + \tilde{Y}$ is also parallel and $(Y + \tilde{Y})|_{\alpha(t)} = v + \tilde{v}$. The same idea works with scaling v.

Some other properties of $P(\alpha)_t^s$ follow easily from its definition as the solution of an ODE. We have semi-group properties $P(\alpha)_t^t = \mathrm{id}$ and $P(\alpha)_s^u \circ P(\alpha)_t^s = P(\alpha)_t^u$. By the uniqueness of the solutions to ODEs, we have that $P(\alpha)_t^s$ is injective, and therefore an isomorphism of vector spaces. And so on.

Conversely, if one has the parallel transport operator for a curve α, the we can recover the covariant derivative in the direction α' through the formula

$$\nabla_{\alpha'(0)}Y = \lim_{h\to 0} \frac{1}{h}\Big[P(\alpha)_h^0 Y|_{\alpha(h)} - Y|_{\alpha(0)}\Big] \in T_{\alpha(0)}M.$$

Exercise 3.35. Prove the above formula. Hint: Take a basis of $T_{\alpha(0)}M$ and parallel transport it along α. As a reward for solving this exercise, you may now use the word *connection* for a covariant derivative.

So intuitively the two approaches, covariant derivatives and parallel transport operators, are equivalent. The reason that it is difficult to start with parallel transport operators is that is tricky to characterise exactly when a set of linear functions between tangent spaces, one for every curve, correspond to a covariant derivative. Note out logic above: if we begin with a covariant

derivative, then we have a parallel transport operator, and taking a limit we can recover the covariant derivative. But if you begin with a arbitrary set of operators, there is no guarantee that the limit will exist. You need to have some type of smooth dependence of $P(\alpha)_t^s$ on t and s. Further, what conditions should you impose on the dependence of $P(\alpha)$ on α such if two curves are tangent at a point, the above limit produces the same result. Hopefully, these questions give you an appreciation of the difficulty involved.

Special mention should go to Appendix B in Sharpe, which does start with the classical idea of rolling a plane (or another space) around on a surface and shows how that gives various modern structures on the manifold.

3.5 Torsion

In this section we discuss a quantity called torsion that is derived from a covariant derivative. There is a relation between the torsion of a connection and the torsion of a space curve, but we will not be explore it in this course[3]. Ultimately we will only be interested in covariant derivatives with zero torsion, so in a sense we are introducing it only to rule it out. Which brings us to the point: how should we motivate the definitions in this section without going deep into theory we will not use? We ask some natural questions and give some reasonable answers.

In euclidean space we have Schwarz' theorem, also known as Clairaut's theorem, that the partial derivatives with respect to different variables commute (for smooth functions among others). This result is embedded in the definition of the Lie bracket, where it was necessary to have the second order terms cancel. In fact sometimes the theorem is expressed as $[\partial_i, \partial_j] = 0$. So naturally we ask this question of the covariant derivative, but the answer is negative in general:

$$\nabla_{\partial_i}\partial_j - \nabla_{\partial_j}\partial_i = \Gamma^k_{ij}\partial_k - \Gamma^k_{ji}\partial_k = \big(\Gamma^k_{ij} - \Gamma^k_{ji}\big)\partial_k.$$

This leads to the following definition

Definition 3.36. *We say that a covariant derivative is* torsion-free *(in some chart) if* $\nabla_{\partial_i}\partial_j - \nabla_{\partial_j}\partial_i = 0$. *Equivalently in terms of Christoffel coefficients, if* $\Gamma^k_{ij} = \Gamma^k_{ji}$ *at every point.*

In this first definition, torsion of a covariant derivative is a measure of the non-commutativity of coordinate vector fields. It seems natural therefore that this should depend on the choice of chart as much as the covariant derivative. But if you have done Exercise 3.31, you may already know that if $\Gamma^k_{ij} = \Gamma^k_{ji}$ at a point in one chart then it also holds at that point in any overlapping chart. We will return to this idea shortly.

Example 3.37. We have $\mathbb{R}^n$ with one chart, and ∇^{euc} from Example 3.19. In Example 3.28 computed that the Christoffel coefficients are all zero. Thus this covariant derivative is torsion-free in this chart.

Example 3.38. In Example 3.29 we gave all the Christoffel coefficients. Observe that they are symmetric in the lower two indices

$$\Gamma^1_{21} = \frac{-2x^2}{\|x\|^2+1} = \Gamma^1_{12}, \qquad \Gamma^2_{21} = \frac{-2x^1}{\|x\|^2+1} = \Gamma^2_{12}.$$

This shows that the covariant derivative $\nabla^\top$ of $\mathbb{S}^2$ is torsion-free on U_N.

We have the expectation that the coordinate vector fields should commute, or that this is a

[3] See the 'American football example' Lee Problem 6-1

desirable property, but we do not have that expectation for general vector fields X, Y. We find

$$\begin{aligned}\nabla_X Y - \nabla_Y X &= X^i \nabla_{\partial_i}(Y^j \partial_j) - Y^j \nabla_{\partial_j}(X^i \partial_i) \\ &= \left(X^i \frac{\partial Y^k}{\partial x^i} + X^i Y^j \Gamma^k_{ij} \right) \partial_k - \left(Y^j \frac{\partial X^k}{\partial x^i} + Y^j X^i \Gamma^k_{ji} \right) \partial_k \\ &= \left(X^i \frac{\partial Y^k}{\partial x^i} - Y^j \frac{\partial X^k}{\partial x^i} \right) \partial_k + X^i Y^j \left(\Gamma^k_{ij} - \Gamma^k_{ji} \right) \partial_k \\ &= [X, Y] + X^i Y^j \left(\Gamma^k_{ij} - \Gamma^k_{ji} \right) \partial_k.\end{aligned}$$

The meaning of this equation is that the 'covariant derivative commutator' of two vector fields is their Lie bracket plus a factor coming from the fact that the coordinate vector fields do not 'covariantly commute'.

Definition 3.39. *Given a covariant derivative ∇, we define the torsion of two vector fields X, Y to be a third vector field*

$$T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y].$$

Remarkably the value of $T(X, Y)$ at any point p only depends on $X|_p, Y|_p$, with the formula

$$T(X, Y) = X^i Y^j T^k_{ij} \partial_k \qquad \text{for} \qquad T^k_{ij} = \Gamma^k_{ij} - \Gamma^k_{ji}.$$

The definition of $T(X, Y)$ is in terms of three vector fields $\nabla_X Y$, $\nabla_Y X$, and $[X, Y]$, so clearly is independent of charts. A covariant derivative is torsion-free if $T^k_{ij} = 0$, and so this too is independent of charts. The second formula is just a rearrangement of the calculation preceding the definition. We say that the second formula is remarkable because although T is defined using derivatives both of which depend on the local behaviour of vector fields, the torsion only depends on the pointwise values of the vector fields. Because the Lie bracket is an antisymmetric function of X, Y, so too is the torsion $T(X, Y) = -T(Y, X)$.

Example 3.40. In this example we show that the torsion of the covariant derivative ∇^L on $\mathbb{S}^3$ from Example 3.27 is non-zero. The trick is to not work with coordinate vector fields, but rather work with left-invariant vector fields. Let $E_1|_p = p\mathbf{i}$ and likewise $E_2|_p = p\mathbf{j}, E_3|_p = p\mathbf{k}$ denote the left-invariant vector fields that are obtained by pushing forward $\mathbf{i}, \mathbf{j}, \mathbf{k} \in T_e\mathbb{S}^3$. We have already noted in Example 3.34 that $\nabla^L_v E_i = 0$ for any vector $v \in T_pM$.

Further at any point $E_1|_p, E_2|_p, E_3|_p$ is a basis for $T_p\mathbb{S}^3$. This means that every vector field X on $\mathbb{S}^3$ can be written as

$$X = X^1 E_1 + X^2 E_2 + X^3 E_3.$$

Thus E_i have similar properties to the coordinate vector basis field, except that they do not come from coordinates. A set of vector fields with this basis property is called a *frame*, but we will not explore this concept in generality. In this frame, the covariant derivative can be reckoned with

$$\nabla^L_X Y = \nabla^L_X \left(Y^j E_j \right) = X(Y^j) E_j + Y^j \nabla^L_X E_j = X(Y^j) E_j.$$

Similarly the Lie bracket simplifies

$$\begin{aligned}[E_i, Y] &= [E_i, Y^j E_j] = E_i(Y^j)E_j + Y^j[E_i, E_j] \\ [X, Y] &= [X^i E_i, Y^j E_j] = X^i[E_i, Y^j E_j] - Y^j E_j(X^i)E_i \\ &= X^i E_i(Y^j)E_j + X^i Y^j[E_i, E_j] - Y^j E_j(X^i)E_i \\ &= X(Y^j)E_j + X^i Y^j[E_i, E_j] - Y(X^i)E_i.\end{aligned}$$

Together this yields

$$\begin{aligned}T^L(X, Y) &= \nabla^L_X Y - \nabla^L_Y X - [X, Y] \\ &= X(Y^j)E_j - Y(X^j)E_j - X(Y^j)E_j - X^i Y^j[E_i, E_j] + Y(X^i)E_i \\ &= -X^i Y^j[E_i, E_j].\end{aligned}$$

Thus the torsion comes down to the Lie brackets of this frame.

For this example we will evaluate $[E_1, E_2]$:

$$\begin{aligned}[E_1, E_2] = [p\mathbf{i}, p\mathbf{j}] &= [-p^1 + p^0\mathbf{i} + p^3\mathbf{j} - p^2\mathbf{k}, -p^2 - p^3\mathbf{i} + p^0\mathbf{j} + p^1\mathbf{k}] \\ &= [-p^1\partial_0 + p^0\partial_1 + p^3\partial_2 - p^2\partial_3, -p^2\partial_0 - p^3\partial_1 + p^0\partial_2 + p^1\partial_3] \\ &= -p^1\partial_2 + p^0\partial_3 + p^3(-\partial_0) - p^2(-\partial_1) - \big[-p^2\partial_1 - p^3(-\partial_0) + p^0(-\partial_3) + p^1\partial_2\big] \\ &= -2p^3\partial_0 + 2p^2\partial_1 - 2p^1\partial_2 + 2p^0\partial_3 = 2E_3.\end{aligned}$$

We can generalise this argument; set $\mathbf{i}_1 = \mathbf{i}, \mathbf{i}_2 = \mathbf{j}, \mathbf{i}_3 = \mathbf{k}$ so that we can use index notation.

$$[E_i, E_j] = [p\partial_i, p\partial_j] = p\mathbf{i}_i\partial_j - p\mathbf{i}_j\partial_i = p(\mathbf{i}_i\mathbf{i}_j - \mathbf{i}_j\mathbf{i}_i).$$

When $i = j$, the quaternions commute and the bracket is zero (as expected). If they are not equal then the quaternions anti-commute. This gives $[E_2, E_3] = 2E_1$ and $[E_3, E_1] = 2E_2$. (There is in fact a close relationship between the Lie bracket of $\mathbb{S}^3$ and the cross product of $\mathbb{R}^3$).

Example 3.41. We can also ask for the torsion of ∇^R on $\mathbb{S}^3$. Of course we could do the same as the previous example, except using a right-invariant frame, and get a similar answer. But to make the two examples comparable, let us compute the torsion of ∇^R using the left-invariant frame E_i.

What changes about the calculation is that $\nabla^R_{E_i} E_j \neq 0$. Instead we must generalise the calculation from Example 3.27:

$$\begin{aligned}E_i(p) &= p\partial_i p = p\mathbf{i}_i, \\ \nabla^R_{E_i} E_j &= \big(E_i(p\mathbf{i}_j p^{-1})\big)\, p = \big(E_i(p)\mathbf{i}_j p^{-1} - p\mathbf{i}_j p^{-1} E_i(p) p^{-1}\big)\, p = p\mathbf{i}_i\mathbf{i}_j - p\mathbf{i}_j\mathbf{i}_i = [E_i, E_j].\end{aligned}$$

The covariant derivative of an arbitrary vector field is

$$\nabla^R_X Y = X(Y^j)E_j + X^i Y^j \nabla^R_{E_i} E_j = X(Y^j)E_j + X^i Y^j[E_i, E_j].$$

Hence

$$\begin{aligned} T^R(X,Y) &= \nabla^R_X Y - \nabla^R_Y X - [X,Y] \\ &= X(Y^j)E_j + X^iY^j[E_i,E_j] - Y(X^j)E_j - X^iY^j[E_j,E_i] \\ &\quad - X(Y^j)E_j - X^iY^j[E_i,E_j] + Y(X^i)E_i \\ &= X^iY^j[E_i,E_j]. \end{aligned}$$

Thus the torsion of ∇^R is the negative of the torsion of ∇^L.

Recall Exercise 3.22 that given one connection we can create another by the addition of a vector valued function $A(X,Y)$. We can ask how the torsion of the new covariant derivative related to the torsion of the original. This follows easily, for $\tilde{\nabla} = \nabla + A$,

$$\begin{aligned} T^{\tilde{\nabla}}(X,Y) &= \tilde{\nabla}_X Y - \tilde{\nabla}_Y X - [X,Y] = \nabla_X Y + A(X,Y) - \nabla_Y X - A(Y,X) - [X,Y] \\ &= T^{\nabla}(X,Y) + A(X,Y) - A(Y,X). \end{aligned}$$

Purely algebraically, for any function of two variables we can split it into a symmetric and antisymmetric parts

$$A(X,Y) = \frac{1}{2}\Big(A(X,Y) + A(Y,X)\Big) + \frac{1}{2}\Big(A(X,Y) - A(Y,X)\Big).$$

If A is already symmetric or antisymmetric, then it is just equal to its symmetric or antisymmetric part respectively and the other part is zero. Thus we can express the relationship of the torsions by the dictum "adding A to a covariant derivative adds twice the antisymmetric part of A to its torsion". In particular, for any covariant derivative, we can *absorb the torsion*. This means we construct a new torsion-free covariant derivative $\tilde{\nabla} := \nabla - \frac{1}{2}T$.

Example 3.42. We have just seen in Examples 3.40 and 3.41 that with respect to the left-invariant fields E_i the covariant derivatives are

$$\nabla^L_{E_i} E_j = 0 \qquad \nabla^R_{E_i} E_j = [E_i,E_j]$$
$$T^L(E_i,E_j) = -[E_i,E_j] \qquad T^R(E_i,E_j) = [E_i,E_j].$$

(Aside: the formula on the right makes it seem as if ∇^R and T^R are equal. They are not in general, only for left-invariant vector fields. Remember: a covariant derivative has the product rule in Y, whereas the torsion is C^∞-linear.)

If we absorb the torsion on these two connections we get the torsion-free connection

$$\nabla^{LC}_{E_i} E_j = \frac{1}{2}[E_i,E_j] = \nabla^L_{E_i} E_j + \frac{1}{2}[E_i,E_j] = \nabla^R_{E_i} E_j - \frac{1}{2}[E_i,E_j].$$

This fits nicely with Corollary 3.23, because ∇^{LC} can also be understood as the average of the left and right covariant derivatives: $\nabla^{LC} = \frac{1}{2}\nabla^L + \frac{1}{2}\nabla^R$. I'll give you one guess what the LC stands for!

We have seen now that for a torsion-free connection that the coordinate vector fields will 'covariant commute' but general vector fields will not.

Definition 3.43. *A smooth family of curves is a function $\alpha_s(t) : (-\varepsilon, \varepsilon) \times (a, b) \to M$. By smooth family we mean that it is smooth in both variables s and t. We typically think of the* main *curves of the family $t \mapsto \alpha_s(t)$ for fixed s. But we also have the* transverse *curves, where we fix t and allow s to vary. We can write $\alpha(s, t)$ to emphasise this duality.*

Therefore we have two vector fields: the tangents in the main direction and the tangents in the transverse direction. Well, this is not completely true as we do not really have vector fields because the curves may cross each other, giving multiple vectors at the same point. (Technically what we have is the pushforwards of two vector fields.) Regardless, for each value of (s, t) it makes sense to ask how the derivative $\partial_s\alpha$ is changing in comparison to $\partial_t\alpha$.

Lemma 3.44 (Mixed Derivatives)**.** *Let ∇ be a torsion-free covariant derivative and $\alpha(s, t) : (-\varepsilon, \varepsilon) \times (a, b) \to M$ a smooth family of curves. Then $\nabla_{\partial_s\alpha}\partial_t\alpha = \nabla_{\partial_t\alpha}\partial_s\alpha$.*

Proof. This is a purely computational proof. In a chart, the tangent vectors are

$$\partial_s\alpha = \frac{\partial\alpha^k}{\partial s}\partial_k, \qquad \partial_t\alpha = \frac{\partial\alpha^k}{\partial t}\partial_k.$$

Then

$$\nabla_{\partial_s\alpha}\partial_t\alpha = \left(\frac{\partial^2\alpha^k}{\partial s\partial t} + \Gamma^k_{ij}\frac{\partial\alpha^i}{\partial s}\frac{\partial\alpha^j}{\partial t}\right)\partial_k,$$
$$\nabla_{\partial_t\alpha}\partial_s\alpha = \left(\frac{\partial^2\alpha^k}{\partial t\partial s} + \Gamma^k_{ij}\frac{\partial\alpha^i}{\partial t}\frac{\partial\alpha^j}{\partial s}\right)\partial_k.$$

By the symmetry of the Christoffel coefficients for torsion-free covariant derivatives, these are equal. □

We should comment about why the expression $\nabla_{\partial_s\alpha}\partial_t\alpha$ is well-defined even though the tangents do not necessarily form a vector field. We know that the direction of ∇ depends only on the pointwise value, so this is no issue. And for $\partial_t\alpha$ we need to know its values along a curve in the direction of $\partial_s\alpha$, but this is exactly the meaning of partial derivative. So understood correctly, these expressions are valid. This is an instance where a fleshed out notion of 'vector field on a curve' would have been clarifying, but hopefully you see that not much has been lost by skipping this concept.

3.6 The Levi-Civita connection

Let us once more return to the thought experiment of walking along the equator $\alpha(t) = (\cos t, \sin t, 0)$ with our stick. We now understand that we are parallel transporting our stick. But consider the vector field $Z(\alpha(t)) = (0, 0, -\cos^2 t)$. To push the metaphor into silliness, it is an telescoping selfie stick that is lengthening and shortening. The vector field Z always points south, but it is not parallel according to definition. If we write $Z = \cos^2 t Y$ for $Z(\alpha(t)) = (0, 0, -1)$ then

$$\nabla^\top_{\alpha'} Z = \nabla^\top_{\alpha'}(\cos^2 t\, Y) = \frac{d}{dt}\Big(\cos^2 t\Big)Y + \cos^2 t\, \nabla^\top_{\alpha'} Y = -2\sin t \cos t\, Y \not\equiv 0.$$

This illustrates the point that parallel is about more than just direction, it also concerns length (which is unlike how we use the term in elementary geometry and linear algebra). Therefore, among the many covariant derivatives that exists on a Riemannian manifold, we are interested in those whose parallel transport preserves length and angle.

Let us now turn this intuition into a definition. Suppose M is a Riemannian manifold with metric g and that ∇ is a connection that preserves the lengths and angles of parallel transport vectors. For any curve α, let X, Y be parallel fields along α with respect to ∇. This means that $g(X, Y)$ is a constant function along α. For all smooth functions a, b, we must have

$$\begin{aligned}\frac{d}{dt} g(aX, bY) &= \frac{d}{dt}\Big(ab\, g(X, Y)\Big) = \frac{da}{dt} b g(X, Y) + a\frac{db}{dt} g(X, Y) + ab\frac{d}{dt} g(X, Y) \\ &= g(a'X, bY) + g(aX, b'Y) + 0.\end{aligned}$$

On the other hand

$$\nabla_{\alpha'}(aX) = a'X + a\nabla_{\alpha'} X = a'X.$$

Therefore we make the definition

Definition 3.45. *A covariant derivative ∇ is called* metric-compatible *or a* metric connection *if for all vector fields X, Y, Z*

$$Z\big(g(X, Y)\big) = g\big(\nabla_Z X, Y\big) + g\big(X, \nabla_Z Y\big).$$

The choice to define this property using a third vector field Z instead of the tangent vector α' is purely a matter of style. The converse of the above argument is immediate: if X, Y are parallel along a curve α then the right hand side is zero and thus $g(X, Y)$ is constant on the curve.

Example 3.46. We can show that the left and right covariant derivatives are compatible with the metric on $\mathbb{S}^3$ coming from $\mathbb{R}^4$. Write vector fields $X = X^i E_i$ and $Y = Y^i E_i$ with respect to the left-invariant basis fields from Example 3.40. By the property of quaternions that $a \cdot b = \operatorname{Re} \bar{a} b$ we see that

$$E_i \cdot E_j = \operatorname{Re} \overline{p\mathbf{i}_i} p \mathbf{i}_j = \operatorname{Re} \bar{\mathbf{i}}_i \bar{p} p \mathbf{i}_j = \operatorname{Re} \bar{\mathbf{i}}_i \mathbf{i}_j = \mathbf{i}_i \cdot \mathbf{i}_j,$$

since $p \in \mathbb{S}^3$ has unit length. In particular it is constant on all of $\mathbb{S}^3$. Additionally, the covariant derivatives of the E_i are zero in every direction. Therefore, similar to the calculation before

the definition, we have

$$\begin{aligned} Z(X\cdot Y) &= Z(X^iY^j\,E_i\cdot E_j) = Z(X^i)Y^j\,E_i\cdot E_j + X^iZ(Y^j)\,E_i\cdot E_j \\ &= \big(Z(X^i)E_i\big)\cdot Y + X\cdot\big(Z(Y^j)E_j\big) \\ &= \big(Z(X^i)E_i + X^i\nabla^L_Z E_i\big)\cdot Y + X\cdot\big(Z(Y^j)E_j + Y^j\nabla^L_Z E_j\big) \\ &= \big(\nabla^L_Z X\big)\cdot Y + X\cdot\big(\nabla^L_Z Y\big). \end{aligned}$$

This shows that ∇^L is metric-compatible.

For ∇^R we can reuse some of this calculation. What changes is that $\nabla^R_Z E_i$ may not be zero. Instead $\nabla^R_Z E_i = Z^k[E_k, E_i]$. We need to prove a version of the cyclic property for the triple product (for vectors in $\mathbb{R}^3$ we have $a\cdot(b\times c) = b\cdot(c\times a)$):

$$\begin{aligned} [E_k,E_i]\cdot E_j + E_i\cdot[E_k,E_j] &= \operatorname{Re}\overline{(\mathbf{i}_k\mathbf{i}_i - \mathbf{i}_i\mathbf{i}_k)}\mathbf{i}_j + \operatorname{Re}\bar{\mathbf{i}}_i(\mathbf{i}_k\mathbf{i}_j - \mathbf{i}_j\mathbf{i}_k) \\ &= \operatorname{Re}(\mathbf{i}_i\mathbf{i}_k\mathbf{i}_j - \mathbf{i}_k\mathbf{i}_i\mathbf{i}_j - \mathbf{i}_i\mathbf{i}_k\mathbf{i}_j + \mathbf{i}_i\mathbf{i}_j\mathbf{i}_k) \\ &= \operatorname{Re}(-\mathbf{i}_k\mathbf{i}_i\mathbf{i}_j + \mathbf{i}_i\mathbf{i}_j\mathbf{i}_k) = 0. \end{aligned}$$

This allows us to write

$$\begin{aligned} Z(X\cdot Y) &= \big(Z(X^i)E_i\big)\cdot Y + X\cdot\big(Z(Y^j)E_j\big) + X^iY^jZ^k([E_k,E_i]\cdot E_j + E_i\cdot[E_k,E_j]) \\ &= \big(Z(X^i)E_i\big)\cdot Y + X\cdot\big(Z(Y^j)E_j\big) + X^i\nabla^R_Z E_i\cdot Y + Y^jX\cdot\nabla^R_Z E_j \\ &= \big(\nabla^R_Z X\big)\cdot Y + X\cdot\big(\nabla^R_Z Y\big). \end{aligned}$$

This proves that ∇^R is also metric-compatible.

The set of metric-compatible covariant derivatives has a affine structure (compare to Corollary 3.23)

Theorem 3.47 (Affineness)**.** *If f $t\in\mathbb{R}$ is a constant and ∇^0,∇^1 are two metric-compatible covariant derivatives, so is $\nabla^t := (1-t)\nabla^0 + t\nabla^1$.*

Proof. We know that ∇^t is a covariant derivative, so it remains to show that it is metric compatible:

$$\begin{aligned} Z(g(X,Y)) &= Z((1-t+t)g(X,Y)) = (1-t)Z(g(X,Y)) + tZ(g(X,Y)) \\ &= (1-t)\left[g\big(\nabla^0_Z X,Y\big) + g\big(X,\nabla^0_Z Y\big)\right] + t\left[g\big(\nabla^1_Z X,Y\big) + g\big(X,\nabla^1_Z Y\big)\right] \\ &= g\big((1-t)\nabla^0_Z X + t\nabla^1_Z, Y\big) + g\big(X,(1-t)\nabla^0_Z Y + t\nabla^1_Z Y\big) \\ &= g\big(\nabla^t_Z X,Y\big) + g\big(X,\nabla^t_Z Y\big). \end{aligned}$$

□

Notice that in the above proof, unlike the proof of Corollary 3.23, it was import that t was a constant so that it could pass through Z. This seems to indicate that metric-compatibility is a rather strong condition. In fact, its almost enough to guarantee uniqueness, but not quite. However, metric-compatibility and torsion-free are enough to do the job. This result is given a rather impressive sounding name, though sometimes it is called a theorem and other times a lemma. We have our cake and eat it too:

Theorem 3.48 (Fundamental Lemma of Riemannian Geometry)**.** *On every Riemannian manifold there exists a unique metric-compatible torsion-free covariant derivative.*

Proof. Our strategy for the proof is as follows. First we will establish the so-called Koszul formula. Uniqueness is then a direct consequence. To prove existence we will show that the Koszul formula defines a torsion-free metric-compatible covariant derivative in every chart. Since we already have uniqueness, we can conclude that these give a well-defined covariant derivative on the whole manifold.

The idea of the Koszul formula is to use the symmetries of the metric and the Lie bracket to get an expression with exact one covariant derivative. Begin with the metric-compatibility property and then use the fact that torsion is zero:

$$\begin{aligned} Z\big(g(X,Y)\big) &= g\big(\nabla_Z X, Y\big) + g\big(X, \nabla_Z Y\big) = g\big(\nabla_Z X, Y\big) + g\big(X, \nabla_Y Z + [Z,Y]\big) \\ &= g\big(\nabla_Z X, Y\big) + g\big(X, \nabla_Y Z\big) + g\big(X, [Z,Y]\big). \end{aligned}$$

Now write this equation two more times with the vector fields permuted

$$\begin{aligned} Y\big(g(Z,X)\big) &= g\big(\nabla_Y Z, X\big) + g\big(Z, \nabla_X Y\big) + g\big(Z, [Y,X]\big) \\ X\big(g(Y,Z)\big) &= g\big(\nabla_X Y, Z\big) + g\big(Y, \nabla_Z X\big) + g\big(Y, [X,Z]\big). \end{aligned}$$

Notice that of the six possible permutations, only $\nabla_Z X$, $\nabla_Y Z$ and $\nabla_X Y$ occur. This is a result of using the torsion free property. Each of the three covariant occurs twice. Now, add any two equations and subtract the other. We will add the second and third and subtract the first, but it's not important which you choose.

$$\begin{aligned} &X\big(g(Y,Z)\big) + Y\big(g(Z,X)\big) - Z\big(g(X,Y)\big) \\ &\qquad = 2g\big(Z, \nabla_X Y\big) + g\big(Z, [Y,X]\big) + g\big(Y, [X,Z]\big) - g\big(X, [Z,Y]\big). \end{aligned}$$

If you like, you can clean this up a little, though the role each of the vector fields play in $g\big(Z, \nabla_X Y\big)$ is different, so there cannot be perfect symmetry in the formula. Here is a version I like:

$$\begin{aligned} 2g\big(\nabla_X Y, Z\big) &= X\big(g(Y,Z)\big) - g\big(X, [Y,Z]\big) \\ &\quad + Y\big(g(X,Z)\big) - g\big(Y, [X,Z]\big) \\ &\quad - Z\big(g(X,Y)\big) + g\big(Z, [X,Y]\big) \end{aligned}$$

This is the Koszul formula. Since the metric is non-degenerate, what it shows is that a metric-compatible torsion-free covariant derivative can be calculated purely in terms of Lie brackets and inner products. Therefore we have established uniqueness.

For existence, it is possible to take the Koszul formula as the definition and directly check all the required properties. It is easier however to first reduce the Koszul formula to an expression in charts. Choose any chart and suppose that $X\partial_i, Y = \partial_j, Z = \partial_k$ are coordinate vector fields. The the Lie brackets are zero. We get

$$2g(\Gamma^l_{\,i\,j}\partial_l, \partial_k) = 2\Gamma^l_{\,i\,j} g_{l\,k} = \partial_i g_{j\,k} + \partial_j g_{i\,k} - \partial_k g_{i\,j}.$$

If we view the left hand side in matrix notation rather than index notation, we see that solve for Γ we need to invert the matrix $G = (g_{i\,j})$. There is a sneaky convention that the components

of the inverse matrix use upper indices $(g^{ij}) = G^{-1}$. With this convention, the fact that these matrices are inverse can be written $g^{ij}g_{jk} = \delta^i_k$. Thus we can write

$$2\Gamma^l_{ij}g_{lk}g^{km} = 2\Gamma^l_{ij}\delta^m_l = 2\Gamma^m_{ij} = g^{km}\big(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}\big). \tag{3.49}$$

So given a metric g, define a covariant derivative on this chart using this formula for the Christoffel coefficients. Exercise 3.30 tells us that this does indeed define a covariant derivative on this chart, but it remains to show that it is metric-compatible and torsion-free. Torsion-free is an easy because the above formula is symmetric in i and j. For metric-compatible we need to write the condition in a chart. Notice that the formula for metric-compatibility is C^∞-linear in Z, so it enough to show it holds for each coordinate basis vector. The following calculation is a set of equivalences:

$$\begin{gathered}
\partial_k\big(g(X,Y)\big) = g\big(\nabla_{\partial_k}X, Y\big) + g\big(X, \nabla_{\partial_k}Y\big) \\
\partial_k\big(X^iY^jg_{ij}\big) = g\big((\partial_kX^i + X^l\Gamma^i_{kl})\partial_i, Y^j\partial_j\big) + g\big(X^i\partial_i, (\partial_kY^j + Y^l\Gamma^j_{kl})\partial_j\big) \\
\partial_kX^iY^jg_{ij} + X^i\partial_kY^jg_{ij} + X^iY^j\partial_kg_{ij} = (\partial_kX^i + X^l\Gamma^i_{kl})Y^jg_{ij} + X^i(\partial_kY^j + Y^l\Gamma^j_{kl})g_{ij} \\
X^iY^j\partial_kg_{ij} = X^l\Gamma^i_{kl}Y^jg_{ij} + X^iY^l\Gamma^j_{kl}g_{ij} \\
\partial_kg_{ij} = \Gamma^l_{ki}g_{lj} + \Gamma^l_{kj}g_{il}.
\end{gathered} \tag{3.50}$$

In other words, a covariant derivative is metric-compatible if and only if its Christoffel coefficients satisfy the above equation. Using the Christoffel coefficients defined through the Koszul formula, we see that this condition is satisfied:

$$2\Gamma^l_{ki}g_{lj} + 2\Gamma^l_{kj}g_{il} = \big(\partial_kg_{ij} + \partial_ig_{kj} - \partial_jg_{ki}\big) + \big(\partial_kg_{ji} + \partial_jg_{ki} - \partial_ig_{kj}\big) = 2\partial_kg_{ij}.$$

Therefore the covariant derivative that we have defined in each chart is metric-compatible and torsion-free. As mentioned at the outset of the proof, it only remains to show that this definition in each chart agrees, but this follows do to uniqueness. □

We celebrate this result with more terminology. It honours the Italian mathematician Tullio Levi-Civita, who developed much of the 'tensor calculus' (covariant, contravariant, indices, etc). His name tricks many students (myself included) into thinking there are two mathematicians Levi and Civita. In response to being asked what he liked best about Italy, Einstein once said "spaghetti and Levi-Civita".

Definition 3.51. *The unique metric-compatible torsion-free covariant derivative on a Riemannian manifold is called the* Levi-Civita connection *or the* Riemannian connection.

Example 3.52. On any open subset of $\mathbb{R}^n$ with the dot product as metric, the Levi-Civita connection is ∇^{euc} from Example 3.19. Because its Christoffel coefficients are identically zero, obviously it is torsion-free and satisfies Equation (3.50) so is metric-compatible.

Example 3.53. We saw in Example 3.46 that both ∇^L and ∇^R were metric-compatible. Therefore their average $\nabla^{LC} = \frac{1}{2}\nabla^L + \frac{1}{2}\nabla^R$ is too, due to Theorem 3.47. Additionally, we proved in Example 3.42 that it is torsion-free. Therefore ∇^{LC} really is the Levi-Civita connection for $\mathbb{S}^3$.

Another obvious example of a Levi-Civita connection would be $\nabla^\top$ on $\mathbb{S}^2$. Instead of proving it for the specific case, instead we generalise the construction to any Riemannian immersed submanifold.

Definition 3.54. *Let $M \to N$ be an Riemannian immersed submanifold. We identify the manifold M with its image under the immersion to simplify the statement. Let ∇^N be the Levi-Civita connection of N. We define the* tangent connection *on M to be the covariant derivative*

$$(\nabla^\top_X Y)|_p = \operatorname*{proj}_{T_pM} \nabla^N_X Y.$$

This is definition extends the previous definition on $\mathbb{S}^2$ because ∇^{euc} is the Levi-Civita connection of $\mathbb{R}^3$.

Theorem 3.55 (Gauss Formula)**.** *The tangent connection is the Levi-Civita connection of M.*

Proof. The proof that it is in fact a covariant derivative is entirely similar to the corresponding statement for the specific case of $\mathbb{S}^2$ in Example 3.24. We check the three properties of a covariant derivative:

$$\begin{aligned}
\nabla^\top_{fX+\tilde{X}} Y &= \operatorname*{proj}_{T_pM} \nabla^N_{fX+\tilde{X}} Y = \operatorname*{proj}_{T_pM} \left(f\nabla^N_X Y + \nabla^N_{\tilde{X}} Y \right) = f\nabla^\top_X Y + \nabla^\top_{\tilde{X}} Y, \\
\nabla^\top_X (Y + \tilde{Y}) &= \operatorname*{proj}_{T_pM} \left(\nabla^N_X Y + \nabla^N_X \tilde{Y} \right) = \nabla^\top_X Y + \nabla^\top_X \tilde{Y}, \\
\nabla^\top_X (fY) &= \operatorname*{proj}_{T_pM} \left(X(f)Y + \nabla^N_X Y \right) = X(f)Y + f\nabla^\top_X Y,
\end{aligned}$$

using that Y is already tangent to M. You might observe that this part of the proof works for any covariant derivative and that we have not yet used the metric-compatibility or torsion-free of ∇^N.

For torsion-free, we need to know that if X, Y are tangent to M that $[X, Y]$ is too, even when we consider them as vector fields on N. To prove this fact requires a proper investigation of submanifolds, and the construction of a special chart on N that aligns with a chart on M. This is beyond the scope of this course, which has tried to avoid manifold theory as much as possible. We have seen an example of this phenomenon though: in Example 3.40 the Lie bracket of the E_i fields was again an E_i field. Assuming this result then,

$$T^\top(X, Y) = \nabla^\top_X Y - \nabla^\top_Y X - [X, Y] = \operatorname*{proj}_{T_pM} \left(\nabla^N_X Y - \nabla^N_Y X - [X, Y] \right) = \operatorname*{proj}_{T_pM} T^N(X, Y)$$

is zero. The generalised statement for arbitrary connections would be that the tangent connection is torsion-free iff the torsion of ∇^N is perpendicular to T_pM at every point of M. Though we hadn't defined it, one could also say iff T^N lies in the normal bundle of M.

Lastly, we need to show that the tangent connection is metric-compatible. This is where we need to use that M is Riemannian immersed, so that the metric on M and the metric on N

agree for tangent vectors to M.

$$\begin{aligned} Z\big(g^M(X,Y)\big) &= Z\big(g^N(X,Y)\big) = g^N\big(\nabla^N_Z X, Y\big) + g^N\big(X, \nabla^N_Z Y\big) \\ &= g^N\big(\operatorname*{proj}_{T_pM} \nabla^N_Z X, Y\big) + g^N\big(X, \operatorname*{proj}_{T_pM} \nabla^N_Z Y\big) \\ &= g^N\big(\nabla^\top_Z X, Y\big) + g^N\big(X, \nabla^\top_Z Y\big) \\ &= g^M\big(\nabla^\top_Z X, Y\big) + g^M\big(X, \nabla^\top_Z Y\big). \end{aligned}$$

To explain the working here a little, for any vector in T_pN we can split it into a part in T_pM and a part perpendicular to T_pM. Because $Y \in T_pM$, the inner product of Y with a vector perpendicular to T_pM is zero. Thus we can go from the first to the second line.

Now we know that $\nabla^\top$ is a metric-compatible torsion-free covariant derivative on M. By the uniqueness in Theorem 3.48, it is the Levi-Civita connection. □

To close the chapter, we revisit the question "why torsion-free"? Our first answer was that it is a natural expectation, based on the commutativity of partial derivatives in the euclidean setting. Our second answer is that torsion-free is a matter of convenience:

- The most common connections, namely the euclidean connection ∇^{euc} and the tangent connection $\nabla^\top$ are torsion-free.
- The Levi-Civita connection is unique and the Koszul formula allows it to be easily calculated.
- For many applications, whether or not a connection has torsion makes no difference. So absorbing the torsion doesn't lead to a loss of generality. We will see this in action in the next chapter.

Chapter 4

Geodesics

This chapter is a continuation of the idea of parallel transport in Section 3.4 . I debated whether to move that section to this chapter, but ultimately decided linking covariant derivatives and parallel transport was necessary as motivation. But perhaps it would be worthwhile to read that section again now.

In this chapter we develop the theory of geodesics, which generalise the notion of a 'straight line'. To put the question provocatively: what does it mean to have a straight line in a curved space? Consider this question for $\mathbb{S}^2$ if somebody asked you to fly an aeroplane in a straight line between two cities. A reasonable definition would be a flight path that did not require steering the plane's control stick. This is the same as the idea of 'walking without turning' that we used previously in our thought experiment. Such paths are called geodesics. On the other hand, this is Riemannian geometry. In euclidean space, the shortest path between two points is a straight line. Perhaps this should be taken as the defining feature of straight lines. It turns out that this *length-minimising* property is also true of geodesics, so that the two possible definitions coincide.

First we will formalise the definition of geodesic. Although after Remark 4.5 we will restrict ourselves to consideration of the Levi-Civita connection, we examine in the first section the mutual dependence of geodesics and connections. Using the length of a curve we define a distance function on a Riemannian manifold as the infimum of the length of all smooth curves. We will show that geodesics are critical points of the length functional, using a proof that is reminiscent of the proof that minimal surfaces are critical for area from Section 1.6. Showing that they are locally length minimising, surprisingly, is significantly more difficult. This leads us to construct special coordinates, so-called normal coordinates, in which the geodesic structure of the Riemannian manifold is a little clearer.

4.1 Straight Lines

Adding to the discussion above, we understand that 'walking without turning' along a curve α means that the tangent of the curve is parallel. The technicality is that α' is not a vector field on the manifold, but this is not a problem because it is a vector field along the curve α and we know that that is sufficient to define its covariant derivative.

Definition 4.1. *A curve α is called a geodesic of a covariant derivative ∇ if*

$$\nabla_{\alpha'}\alpha' = 0.$$

In a chart we may write $\alpha = (\alpha^i)$ and use the Christoffel coefficients to describe the covariant derivative:

$$0 = \nabla_{\alpha'}\alpha' = \frac{d\alpha^i}{dt}\frac{\partial}{\partial x^i}\left(\frac{d\alpha^k}{dt}\partial_k\right) + \frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt}\Gamma^k_{ij}\partial_k = \left[\frac{d^2\alpha^k}{dt^2} + \Gamma^k_{ij}\frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt}\right]\partial_k.$$

This is generally know as the geodesic equation. It is of course the parallel transport equation (3.33) with $Y = \alpha'$. But since the vector field in this case is linked the the curve, we now have a second-order nonlinear system of ODEs. We can use the standard trick to write this as a system of first order ODEs:

$$\frac{d\alpha^k}{dt} = v^k(t), \qquad \frac{dv^k}{dt} = -\Gamma^k_{ij}(\alpha(t))\, v^i(t)v^j(t).$$

From this we conclude the local existence and uniqueness of geodesics in a neighbourhood of every point and in every direction. Specifically, given a point $p \in M$ and a direction $v \in T_pM$ existence implies that there is a curve $\gamma_{p,v} : (a, b) \to M$ with $\gamma_v(0) = 0$ and $\gamma_v'(0) = v$. We will drop the p when it is unambiguous from context. We assume that $\gamma_{p,v} : (a, b) \to M$ is maximal in the sense that any other solution is a restriction of the γ_V to some smaller domain $(a', b') \subset (a, b)$. This assumption may always be achieved, because uniqueness makes it possible to glue together any two solutions into a 'longer' one. Related to this is the observation that if $\tilde{p} = \gamma_{p,w}(t_0)$ and $\tilde{w} = \gamma'_{p,w}(t_0)$, then

$$\gamma_{\tilde{p},\tilde{w}}(t) = \gamma_{p,w}(t + t_0).$$

Notice that the solution is a parameterised curve. But our intuition tells us that a straight line is about direction, not speed. The initial speed determines the speed at every point on the geodesic, so in some sense geodesics are constant speed. We will give a precise statement and proof for metric-compatible connections in Theorem 4.6. In fact, if one has a curve with the property that $\nabla_{\alpha'}\alpha' = f(t)\alpha'(t)$, then it is possible to reparameterise α to be a geodesic. Consider $\beta(t) = \alpha(g(t))$. Then $\beta'(t) = g'(t)\alpha'(g(t))$ and $\beta''(t) = g''(t)\alpha'(g(t)) + g'(t)^2\alpha''(g(t))$ and the geodesic equation reads

$$\begin{aligned}
\frac{d^2\beta^k}{dt^2} + \Gamma^k_{ij}\frac{d\beta^i}{dt}\frac{d\beta^j}{dt} &= g''(t)(\alpha^k)'(g(t)) + g'(t)^2(\alpha^k)''(g(t)) + \Gamma^k_{ij}g'(t)(\alpha^i)'(g(t))g'(t)(\alpha^j)'(g(t)) \\
&= g''(t)(\alpha^k)'(g(t)) + g'(t)^2\left[(\alpha^k)''(g(t)) + \Gamma^k_{ij}(\alpha^i)'(g(t))(\alpha^j)'(g(t))\right] \\
&= g''(t)(\alpha^k)'(g(t)) + g'(t)^2\left[f(g(t))(\alpha^k)'(g(t))\right] \\
&= \left[g''(t) + g'(t)^2 f(g(t))\right](\alpha^k)'(g(t)).
\end{aligned}$$

Now set the bracket to zero and 'just' solve this nasty looking ODE for g to get the necessary reparameterisation.

Some reparameterisations are geodesic preserving. The following lemma tells us how geodesics in the same direction with different initial speeds are related. It says that they have the same image and differ only by a constant rescaling factor. For this reason, geodesics that only differ by speed are often conflated with one another.

Lemma 4.2 (Rescaling Lemma)**.** *Let $w \in T_pM$ and $c \in \mathbb{R} \setminus \{0\}$. Whenever one expression is defined, so is the other and they are equal:*

$$\gamma_w(ct) = \gamma_{cw}(t).$$

Proof. We use the first-order version of the geodesic equation, because it makes the idea a little clearer. As per definition, γ_w is the geodesic with an initial direction of w. Let $v = \gamma_w'$. It obeys

$$\frac{d\gamma_w^k}{dt} = v^k(t), \qquad \frac{dv^k}{dt} = -\Gamma^k_{ij}(\gamma_w(t))\, v^i(t)v^j(t), \qquad \gamma_w(0) = p, \qquad v(0) = w.$$

Now consider the reparameterised curve $\alpha(t) = \gamma_w(ct)$. The starting point has not changed: $\alpha(0) = \gamma_w(0) = p$. But the velocity has changed:

$$\frac{d\alpha(t)}{dt} = \frac{d\gamma_w^k(ct)}{dt} = c\frac{d\gamma_w^k}{dt}(ct) = cv(ct) =: u(t).$$

In particular $u(0) = cv(0) = cw$. Finally, we see for the second ODE

$$\frac{du^k}{dt} = c\frac{dv^k(ct)}{dt} = c^2\frac{dv^k}{dt}(ct) = -c^2\Gamma^k_{ij}(\gamma_w(ct))\, v^i(ct)v^j(ct) = -\Gamma^k_{ij}(\alpha(t))\, u^i(t)u^j(t).$$

Thus we see that (α, u) is a solution of the same ODE and IVP that $\gamma_{cw}(t)$ solves. By uniqueness and maximality of γ_{cw}, we conclude that α is the restriction of γ_{cw}. But we can also run this argument in the reverse direction and conclude that $\tilde{\alpha}(t) = \gamma_{cw}(c^{-1}t)$ is a restriction of γ_w. The conclusion must be that one exists if and only if the other does, and they are equal. □

A particular case is for $c = -1$. The lemma says $\gamma_{-w}(t) = \gamma_w(-t)$: the geodesic in the reverse direction is the same as walking backwards. If we choose a time t_0 and set $\tilde{p} = \gamma(t_0)$ and $\tilde{w} = \gamma'_{p,w}(t_0)$, then we have

$$\gamma_{\tilde{p},-\tilde{w}}(t_0) = \gamma_{\tilde{p},\tilde{w}}(-t_0) = \gamma_{p,w}(-t_0 + t_0) = p.$$

This says that if you walk along a geodesic for a certain amount of time t_0, turn around and walk back for the same amount of time, then you end back where you started. These properties might seem obvious, but it is important to question whether our intuition carry over to the general setting.

Notice that we have excluded the case $c = 0$ in the above lemma, although the result is still valid.[1] In this case we have the claim $\gamma_0(t) = \gamma_w(0) = p$ for all $t \in \mathbb{R}$ and $w \in T_pM$. This degenerate example is none-the-less important to include. These are called *constant* geodesics, because as a function $\mathbb{R} \to M$ they are constant. The interpretation is that walking in a straight line with initial speed zero means standing still.

Example 4.3. Consider the helicoid, which was our primary example in Chapter 1, with the inherited metric from $\mathbb{R}^3$ and the tangent connection. We consider again in this example two (sets of) curves: radial lines and helices. First there are radial curves $\alpha(t) = \Phi(t, v) = (t\cos v, t\sin v, bv)$ for some $v \in \mathbb{R}$. The tangent connection is the projection of the derivational derivative in $\mathbb{R}^3$ to the tangent plane of the helicoid. The directional derivative in the direction α' is the derivative with respect to t. Hence $\nabla^{\text{euc}}_{\alpha'}\alpha' = \alpha'' = 0$, because α is linear in t. Hence radial lines are geodesics of the helicoid.

[1] This is an oversight in the proof of Lee Lemma 5.8, which is not on the list of errata.

On the other hand we have the helices $\beta(t) = \Phi(u,t) = (u\cos t, u\sin t, bt)$ for some $u \in \mathbb{R}$. As above $\nabla^{\text{euc}}_{\beta'}\beta' = \beta''$ but this time it is not zero. Now we reuse our previous calculations. As remarked upon in Example 1.19, β'' already lies in the tangent plane. Hence

$$\nabla^{\top}_{\beta'}\beta' = \operatorname*{proj}_{T_\beta\Sigma}\beta'' = \beta'' = (-u\cos t, -u\sin t, 0) \neq 0.$$

This shows that the helix is not a geodesic for helicoid, unless $u = 0$ (in which case it is the central axis of the helicoid).

We promised to discuss how geodesics depend on the connection. Following Lemma 3.21 we consider the connections ∇ and $\tilde{\nabla} = \nabla + A$. In a chart we may write $A(\partial_i, \partial_j) = A^k_{ij}\partial_k$ and since A is C^∞-bilinear these functions completely determine A. Moreover $\tilde{\Gamma}^k_{ij} = \Gamma^k_{ij} + A^k_{ij}$. Let α be a geodesics of ∇. Then the geodesic equation for $\tilde{\nabla}$ reads

$$\frac{d^2\alpha^k}{dt^2} + \left(\Gamma^k_{ij} + A^k_{ij}\right)\frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt} = \frac{d^2\alpha^k}{dt^2} + \Gamma^k_{ij}\frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt} + A^k_{ij}\frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt} = A^k_{ij}\frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt}.$$

Thus α is also a geodesic of $\tilde{\nabla}$ if and only if this quantity is zero. What is obscured by index notation is the symmetry here. Using a relabelling of summation indices

$$2A^k_{ij}\frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt} = A^k_{ij}\frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt} + A^k_{ji}\frac{d\alpha^j}{dt}\frac{d\alpha^i}{dt} = \left(A^k_{ij} + A^k_{ji}\right)\frac{d\alpha^i}{dt}\frac{d\alpha^j}{dt}.$$

This can of course be zero for some curve α. If A is antisymmetric, $A(X,Y) = -A(X,Y)$, then this quantity is zero for all curves, and in particular the two connections have the same geodesics. Conversely, we know that every vector $w \in T_pM$ at every point is the tangent vector to some geodesic of ∇. If the two share geodesics, then this forces A to be antisymmetric. Therefore we have proved

Theorem 4.4 (Geodesic Agreement)**.** *Two covariant derivatives have exactly the same geodesics if and only if they differ by an antisymmetric A.*

In particular, if we absorb the torsion of a connection $\tilde{\nabla} = \nabla - \frac{1}{2}T$, then this does not change the geodesics. This justifies our comment at the end of Chapter 3 that torsion is often unimportant.

Remark 4.5. For the remainder of the chapter we will only consider the Levi-Civita connection.

We have talked several times in this section about geodesics being constant speed, but we meant this in an informal sense. The speed of a curve only really makes sense in the context of Riemannian geometry, where it is $\|\alpha'\|_g$ the norm of the tangent vector α' with respect to the metric g.

Theorem 4.6 (Constant Speed)**.** *Let M be a Riemannian manifold and ∇ a metric-compatible connection. Then the geodesics of ∇ have constant speed.*

Proof. To remove the square root from the norm, observe that a continuous function is constant if and only if its square is constant. The result now follows from the definition of metric-compatibility:

$$\frac{d}{dt}g(\alpha',\alpha') = g(\nabla_{\alpha'}\alpha',\alpha') + g(\alpha',\nabla_{\alpha'}\alpha') = 0. \qquad \square$$

Finally, what can be said about geodesics in an immersed manifold? For this question to be sensible, we should use the tangent connection on M from Definition 3.54. To recap, we have a Riemannian manifold N with the Levi-Civita connection and M is Riemannian immersed in N. The tangent connection on M is the projection of ∇^N to the tangent space of M. A curve α on M is a geodesic iff

$$0 = \nabla^\top_{\alpha'}\alpha' = \operatorname*{proj}_{T_pM}\left(\nabla^N_{\alpha'}\alpha'\right).$$

Because $\nabla_{\alpha'}\alpha'$ is a second-order derivative of α we interpret it as a type of acceleration. The above equation says that a curve is a geodesic of M when its acceleration in N is always perpendicular to M.

We can relate this back to normal curvature in Section 1.5. In that situation we have $N = \mathbb{R}^3$ and the connection is just ordinary directional derivatives ∇^{euc}. Therefore $\nabla^N_{\alpha'}\alpha' = \alpha''$. The acceleration α'' of a curve in $\mathbb{R}^3$ has components in the T and N directions, Equation (1.24). If we project this onto the tangent plane then we get

$$\nabla^\top_{\alpha'}\alpha' = \|\alpha'\|\left(\frac{d\|\alpha'\|}{ds}T + \|\alpha'\|\kappa \operatorname*{proj}_{T_pM} N\right).$$

Hence we see that α is a geodesic of the surface M if and only if α is parameterised with constant speed and N, the normal of the curve, is perpendicular to M. Since any regular curve can be reparameterised by arc-length, the first condition is only a technical point. If N is perpendicular to M this means that the angle between N and the surface normal ν is zero. In other words, the normal curvature is equal to the curvature of α. Traditionally, one defines the *geodesic curvature* of α as

$$\kappa_g = \kappa\|\operatorname*{proj}_{T_pM} N\| = \kappa\sin\theta.$$

This is a measure of how far an arc-length parameterised curve is from being a geodesic. Alternatively, the observation $\kappa_n^2 + \kappa_g^2 = \kappa^2$ (since the normal and geodesic curvatures are projections of the curvature vector) leads us to say a curve is a geodesic if its curvature is entirely normal.

Example 4.7. Consider $\mathbb{S}^2$. Consider a geodesic α and suppose without loss of generality that α is parameterised by arc-length. From the above discussion, a geodesics α must have α'' in the normal direction to the sphere, $N \cdot \nu = \pm 1$. For the sphere $\nu = \alpha$. Therefore $\nu' = \alpha' = T$. On the other hand $N' = -\kappa T + \tau B$, using the Frenet equations, and we can conclude that $\tau \equiv 0$. This proves that the geodesics of the sphere lie in the plane containing the normal and their tangent vector. These are exactly the *great circles*.

4.2 The Hyperbolic Plane

So far we have seen the example of $\mathbb{S}^2$ and $\mathbb{R}^2$. We have the sense that these spaces are special. There is a another important two-dimensional Riemannian manifold: the *hyperbolic plane* $\mathbb{H}^2$. Unlike the sphere, a theorem of Hilbert proves that the hyperbolic plane cannot be isometrically immersed into $\mathbb{R}^3$[2]. Therefore we really need to use the tools of manifolds and charts to understand this space, there can be no resorting to geometric tricks in euclidean space.

This space was first discovered in connection to the 'parallel postulate' of Euclid. Euclid gave five postulates[3], which we would call axioms,

- α' Let it have been postulated to draw a straight-line from any point to any point.
- β' And to produce a finite straight-line continuously in a straight-line.
- γ' And to draw a circle with any center and radius.
- δ' And that all right-angles are equal to one another.
- ε' And that if a straight-line falling across two (other) straight-lines makes internal angles on the same side (of itself whose sum is) less than two right-angles, then the two (other) straight-lines, being produced to infinity, meet on that side (of the original straight-line) that the (sum of the internal angles) is less than two right-angles (and do not meet on the other side).

Clearly one of these is not like the others. The fifth postulate is called the parallel postulate, because it is equivalent to Playfair's axiom:

> There is at exactly one line that can be drawn parallel to another given one through an external point.

There are several interesting geometries that come from replacing this axiom. If there are no parallel lines then we get projective geometry (one intersection) or spherical geometry (two intersections). If there are more than one parallel line then we get hyperbolic geometry.

For our purposes we will introduce the hyperbolic plane by fiat. Like the euclidean plane it can be covered by a single coordinate chart. There are several 'models' of the hyperbolic plane, but we will use the 'half-plane' model where the geodesics have the easiest formulas.

Definition 4.8. *The hyperbolic plane is the following manifold: Let* $\mathcal{I} = \{0\}$,

$$U_0 = \mathbb{H}^2 = \big\{(x, y) \in \mathbb{R}^2 \mid y > 0\big\},$$

and $\varphi_{0\,0} = \mathrm{id}$, *so that* $(2, \mathcal{I}, \{U_0\}, \{\varphi_{0\,0}\})$ *is its atlas. Additionally, it is a Riemannian manifold with the metric*

$$g_{i\,j}(x, y) = \begin{pmatrix} y^{-2} & 0 \\ 0 & y^{-2} \end{pmatrix}$$

given in the U_0 *chart.*

[2] It is unknown whether it can be isometrically immersed into $\mathbb{R}^4$, but it can be into $\mathbb{R}^5$

[3] Here I am using the translation of Richard Fitzpatrick, based on the authoritative Greek edition of J L Heiberg. A beautiful version designed for teaching can be found here, but it changes the numbering.

The first common misconception to address is the 'boundary' at $y = 0$. As a Riemannian manifold, there is no boundary. If you lived in the hyperbolic plane, as you tried to approach $y = 0$ at constant speed you would find that the coordinates were changing at an every decreasing rate. If we explain this using vectors in coordinates, for y small, a unit-length vector has very small coefficients, and therefore moving at unit-speed makes only a small change in coordinates. Conversely, for large y, a unit-length vector has large coefficients. This is analogous to looking at the world in the Mercator map, where a plane on the equator changes its longitude far less than a plane near the poles even at the same speed.

Let us compute the Christoffel coefficients for the Levi-Civita connection. Only two derivatives of the metric are non-zero

$$\partial_2 g_{1\,1} = -2y^{-3}, \qquad \partial_2 g_{2\,2} = -2y^{-3}$$

and the inverse matrix is

$$g^{i\,j} = \begin{pmatrix} y^2 & 0 \\ 0 & y^2 \end{pmatrix}.$$

Using Equation (3.49)

$$\begin{aligned} 2\Gamma^k_{i\,j} &= y^2\big(\partial_i g_{j\,k} + \partial_j g_{i\,k} - \partial_k g_{i\,j}\big) \\ \Gamma^2_{2\,2} &= \Gamma^1_{1\,2} = \Gamma^1_{2\,1} = -y^{-1} \\ \Gamma^2_{1\,1} &= y^{-1}, \end{aligned}$$

and the other four are zero.

This means that we can try to solve the geodesic equation. It is certainly possible to solve it directly, but we will solve it in a special case and then use isometries to obtain the full solution, as this is geometrically more interesting. The special case is the following: let $p = (0, 1)$ the the starting point of the geodesic and $v = \partial_2$ be its initial direction. If the geodesic is $\alpha(t) = (x(t), y(t))$ then the geodesic formula is

$$\begin{aligned} 0 &= x'' + \Gamma^1_{1\,2} x' y' + \Gamma^1_{2\,1} y' x' = x'' - 2y^{-1} x' y' \\ 0 &= y'' + \Gamma^2_{1\,1} x' x' + \Gamma^2_{2\,2} y' y' = y'' + y^{-1}(x')^2 - y^{-1}(y')^2. \end{aligned}$$

Notice that if (x, y) is a solution to these ODEs, so too is $(-x, y)$. Moreover, $-x(0) = -0 = 0$ and $-x'(0) = -0 = 0$, so they have the same initial condition. Therefore they must be equal, in other words $x(t) = -x(t)$, which implies $x(t) \equiv 0$. The second ODE now simplifies to

$$\begin{aligned} 0 &= y'' - y^{-1}(y')^2 \\ 0 &= \frac{y''}{y'} - \frac{y'}{y} \\ C &= \ln y' - \ln y. \end{aligned}$$

The initial conditions $y(0) = y'(0) = 1$ force $C = 0$. We continue

$$\begin{aligned} \frac{y'}{y} &= 1 \\ y &= Ae^t = e^t, \end{aligned}$$

again using the initial conditions. Thus the unique geodesic through $(0,1)$ with initial vector ∂_2 is the vertical line $\gamma(t) = (0, e^t)$.

Already in this argument we already saw a glimpse to the method we will use to find all other geodesics. We used the *reflection* $R : (x,y) \mapsto (-x,y)$. This preserved the geodesic equations because this is in fact an isometry of $\mathbb{H}^2$. After we strip away the terminology of Definitions 3.4 and 3.8, what we see is that an isometry is a diffeomorphism, in particular a bijective map, such that the pushforward of vector doesn't change the inner product. The pushforward of $v \in T_{(x,y)}\mathbb{H}^2$ by reflection is

$$TR(v^1\partial_1\big|_p + v^2\partial_2\big|_p) = -v^1\partial_1\big|_q + v^2\partial_2\big|_q$$

for $p = (x,y)$ and $q = (-x,y)$, and

$$\begin{aligned} g(TR(v), TR(w)) &= g(-v^1\partial_1\big|_q + v^2\partial_2\big|_q, -w^1\partial_1\big|_q + w^2\partial_2\big|_q) \\ &= v^1w^1g_{1\,1}(q) + v^2w^2g_{2\,2}(q) \\ &= v^1w^1g_{1\,1}(p) + v^2w^2g_{2\,2}(p) \\ &= g(v,w), \end{aligned}$$

because the components of the metric are independent of x.

Another obvious isometry is *horizontal translation* $(x,y) \mapsto (x+a, y)$. From this we can conclude that the geodesic through $(a,1)$ with initial direction ∂_2 is $\gamma(t) = (a, e^t)$. These are simply all the vertical lines.

The translation $(x,y) \mapsto (x, y+a)$ is not an isometry. For one thing, it is not even a well-defined map on $\mathbb{H}^2$, which only has the points of the upper half-plane. But even between points where it is well defined, we see that the metric is changed because $g_{1\,1}(x, y+a) \neq g_{1\,1}(x,y)$.

It turns out that simple scaling of the points $(x,y) \mapsto (rx, ry)$ is an isometry. This is called *dilation*. We compute

$$\begin{aligned} g(TD_r(v), TD_r(w)) &= g(rv^1\partial_1\big|_q + rv^2\partial_2\big|_q, rw^1\partial_1\big|_q + rw^2\partial_2\big|_q) \\ &= r^2v^1w^1g_{1\,1}(q) + r^2v^2w^2g_{2\,2}(q) \\ &= r^2v^1w^1(ry)^{-2} + r^2v^2w^2(ry)^{-2} \\ &= v^1w^1g_{1\,1}(p) + v^2w^2g_{2\,2}(p) \\ &= g(v,w). \end{aligned}$$

Unfortunately, this does tell us any new geodesics, because it takes vertical lines to vertical lines.

We need a different idea. Notice that because the matrix of the metric is a scalar of the euclidean metric, hyperbolic angles and euclidean angles are equal:

$$\cos\theta_{\text{hyp}} = \frac{g(v,w)}{\|v\|_g\|w\|_g} = \frac{y^{-2}\,v\cdot w}{y^{-1}\|v\|\,y^{-1}\|w\|} = \cos\theta_{\text{euc}}.$$

Therefore we should look for transformations of the upper half-plane that preserve euclidean angles (conformal). We saw in Example 3.12 that stereographic projection is a conformal map. Therefore the transition between stereographic projections for the north and south pole, *circle*

inversion, is conformal. It restricts to give a bijective map on the upper half-plane $S(x,y) = (x^2+y^2)^{-1}(x,y)$, because $(0,0) \notin \mathbb{H}^2$. We have already computed the tangent map of this transformation in Example 2.18. We see that, for $p=(x,y)$, $q=S(x,y)$,

$$\begin{aligned}
&g\left(TS(\partial_1|_p), TS(\partial_1|_p)\right)\\
&= g\left(\|p\|^{-4}[y^2-x^2]\partial_1|_q + \|p\|^{-4}[-2xy]\partial_2|_q, \|p\|^{-4}[y^2-x^2]\partial_1|_q + \|p\|^{-4}[-2xy]\partial_2|_q\right)\\
&= \|p\|^{-8}[y^2-x^2]^2 g\left(\partial_1|_q, \partial_1|_q\right) + \|p\|^{-8}[-2xy]^2 g\left(\partial_2|_q, \partial_2|_q\right)\\
&= \|p\|^{-8}\left([y^2-x^2]^2 + [-2xy]^2\right)\left(\|p\|^{-2}y\right)^{-2}\\
&= \|p\|^{-4}\left(y^4 - 2y^2x^2 + x^4 + 4x^2y^2\right)y^{-2}\\
&= y^{-2} = g(\partial_1|_p, \partial_1|_p).
\end{aligned}$$

The calculation for $g\left(TS(\partial_2|_p), TS(\partial_2|_p)\right)$ is exactly the same. That leaves

$$\begin{aligned}
&g\left(TS(\partial_1|_p), TS(\partial_2|_p)\right)\\
&= g\left(\|p\|^{-4}[y^2-x^2]\partial_1|_q + \|p\|^{-4}[-2xy]\partial_2|_q, +\|p\|^{-4}[-2xy]\partial_1|_q + \|p\|^{-4}[x^2-y^2]\partial_2|_q\right)\\
&= \|p\|^{-4}[y^2-x^2]\|p\|^{-4}[-2xy]g\left(\partial_1|_q, \partial_1|_q\right) + \|p\|^{-4}[-2xy]\|p\|^{-4}[x^2-y^2]g\left(\partial_2|_q, \partial_2|_q\right)\\
&= 0 = g(\partial_1|_p, \partial_2|_p).
\end{aligned}$$

Therefore $S(x,y) = (x^2+y^2)^{-1}(x,y)$ is an isometry, not just conformal.

Remarkably circle inversion and the horizontal translations are all the isometries we need to get all the geodesics. By applying S to the geodesic $\gamma(t) = (a, e^t)$ we get the geodesic

$$\left(\frac{a}{a^2+e^{2t}}, \frac{e^t}{a^2+e^{2t}}\right).$$

Looking on a graphing tool we see that they are semicircles centered on the x-axis. This is easy to verify algebraically

$$\left(\frac{a}{a^2+e^{2t}} - \frac{1}{2a}\right)^2 + \left(\frac{e^t}{a^2+e^{2t}}\right)^2 = \frac{(2a^2-a^2-e^{2t})^2 + 4a^2e^{2t}}{4a^2(a^2+e^{2t})^2} = \frac{a^4+2a^2e^{2t}+e^{4t}}{4a^2(a^2+e^{2t})^2} = \frac{1}{4a^2}.$$

In particular we see that we have semicircles of every radius. By horizontal translation then, the set of geodesics includes every vertical line as well as every semicircle centered on the x-axis (from now on we will just say semicircle, leaving the centering implicit).

But in fact this accounts for every unit-speed geodesic. Choose any point p and any direction v with $\|v\|_g$. If v is $\pm\partial_2$, then the desired geodesic is the vertical line. For any other v, by straightedge and compass (for coolness factor) one can construct a semicircle through that point with the direction as tangent. This must be the unique geodesic. For non-unit-speed geodesics, one can simply rescale t, using Lemma 4.2.

Exercise 4.9. Find all geodesics by solving the geodesic equation directly.

These isometries are not only transitive on the set of geodesics, but also on the set of points. This is easily achieved by dilating $(0,1)$ to $(0,y)$ and then a horizontal translation to (x,y). Every

point can therefore be mapped to any other by first bringing it to $(0,1)$ and then sending it on its way. Spaces with a transitive set of isometries are called *homogeneous*. But the hyperbolic plane is also *isotropic*, it looks the same in every direction. This means that there is a full set of isometric rotations. Begin with $\partial_2\big|_{(0,1)}$. Dilate and translate it to any point on a different vertical line as above. We know that this vertical line is transformed by circle inversion to a semicircle. A semicircle has every possible tangent direction except $\pm\partial_2$. You can now translate and dilate it back to $(0,1)$. This give every direction except $-\partial_2$. But this can be achieved through two $\pi/2$ rotations.

Exercise 4.10. Argue that there is exactly one isometry that fixes $(0,1)$ and rotates its tangent space by a given angle.

Exercise 4.11. Prove that a homogeneous space is isotropic at one point if and only if it is isotropic at every point.

Now that we have determined the geodesics, we can comment on the similarities and differences to other geometries. The first observation is that between any two points there is a unique geodesic. Secondly any two geodesics can intersect at most once. Both of these properties are similar to the euclidean plane but are in contrast to spherical geometry, where antipodal points have infinitely many geodesics between them and every pair of geodesics intersects twice.

We can also consider the parallel postulate. Choose a geodesic and a point p external to it. We know that there is an isometry that will transform the geodesic into the y-axis, so we assume this without loss of generality. What are the geodesics through the point that do not cross the y-axis? Not only is there the vertical line through p, there are infinitely many semicircles. In particular, there is one semicircle through p which is tangent to the y-axis. This is called a *limiting parallel*. The vertical line through p is also a limiting parallel, based on our earlier explanation that as points increase in y value for constant x value, the distance between them shrinks.

A *triangle* in the hyperbolic plane is of course a shape with three geodesic sides, none of which are parallel. Just as for euclidean geometry, we can try to classify triangles up to isometry, which is called congruence in elementary geometry. By isometry, we may bring any edge of the triangle to the y-axis, so we assume that one vertex is at $(0,1)$ and the other $(0,b)$. Let α be the angle at $(0,1)$ and β the angle at $(0,b)$. We see that this information already uniquely determines a triangle, thus we have the 'angle-side-angle' rule of triangle congruence.

But more is true. There are restrictions on the angles for two other sides to intersect be able to intersect, namely $\alpha+\beta<\pi$, just as for euclidean space. But unlike the classical situation, even if $\alpha+\beta<\pi$ this is no guarantee that the two sides meet. For any α,β, there is some $B>1$ such that for $b=B$ the sides are limiting parallels. As b ranges from B down to 1, the third angle ranges from 0 (limiting parallel) to $\pi-\alpha-\beta$ (all three corners very close to $(0,1)$, so we can approximate the circles by their tangents). Therefore the sum of angles in a hyperbolic triangle is always less that π. Also note that the angle increases monotonically. This means there is a 1-to-1 correspondence between b and the third angle. In other words, two hyperbolic triangle are congruent if and only if they have corresponding angles: the 'angle-angle-angle' rule!

Exercise 4.12. Admire this Escher drawing.

We close this section with a useful calculation method. It is often profitable to use complex numbers $x+\iota y$ to describe the points of $\mathbb{H}^2$. The chief advantage is the ease of writing isometries. Clearly horizontal translation and dilations are $z \mapsto z + a$ and $z \mapsto rz$ for $a, r \in \mathbb{R}$ respectively. But inversion in the unit circle is $z \mapsto \bar{z}^{-1}$ and reflection in the y-axis is $z \mapsto -\bar{z}$. As you can see by the presence of complex conjugation, these later two isometries are orientation reversing; they are both reflections after all. Therefore it is common to combine them to an isometry $z \mapsto -z^{-1}$. This along with translation and dilation generate all orientation preserving isometries. A general isometry is therefore

$$z \mapsto \frac{az + b}{cz + d}$$

for $a, b, c, d \in \mathbb{R}$ with $ad-bc = 1$. Transformations of this form are called *Möbius transformations*. The set of isometries is three-dimensional so it was necessary to normalise the four constants, and using the 'determinant' makes the formula for the inverse transform simple.

4.3 Length and Distance

In the above discussion of hyperbolic triangle, we were careful not to speak of the length of the sides because we simply have not yet defined length on a Riemannian manifold. Let us remedy that situation. In euclidean space, we had a distance function already and in Theorem 1.5 were able to show that the length of a path as in Definition 1.3 was the integral of its speed. In the general case, we have no prior distance function, but the Riemannian metric does give allows us to determine the speed. Therefore we 'reverse' these theorems:

Definition 4.13. *In a Riemannian manifold M with metric g, the (Riemannian)* length *of a smooth path $\alpha : [a, b] \to M$ is defined to be*

$$L(\alpha) = \int_a^b \|\alpha'(t)\|_g \, dt.$$

The distance d^g *between two points is defined to be the infimum of the lengths of all smooth paths connecting them.*

Example 4.14. With this definition of length we can calculate the lengths of geodesics in the hyperbolic plane. Because we obtained the geodesics by isometries, which preserve the metric and therefore the length, it is sufficient to calculate the length of the vertical geodesic. It has the constant speed parameterisation

$$\gamma(t) = (0, e^t).$$

Therefore

$$\begin{aligned} \gamma'(t) &= (0, e^t), \\ \|\gamma'(t)\|_g^2 &= \|(0, e^t)\|_g^2 = \frac{1}{(e^t)^2}\left[0^2 + (e^t)^2\right] = 1, \\ L(\gamma|_{[a,b]}) &= \int_a^b \|\gamma'(t)\|_g \, dt = \int_a^b \|\gamma'(t)\|_g \, dt = b - a. \end{aligned}$$

To put this more geometrically, the length of the geodesic connecting $(0, y_1)$ and $(0, y_2)$ is $|\ln y_2 - \ln y_1| = \ln |y_2/y_1|$.

There are (at least) two equivalent definitions of the distance function on a Riemannian manifold in the literature, with the difference being over which set of paths the infimum is taken. We have chosen 'smooth curves'. Perhaps more common is to choose 'piecewise smooth curves'. A curve is piecewise smooth if it is continuous and the set of non-smooth points is finite. These non-smooth points are called corners. Smooth functions are clearly nicer to work with than piecewise smooth function, if you work with the latter you are forever splitting the curve into its smooth interval and you must provide correction terms for the corners. The advantage of the piecewise approach is clear if you try to prove the triangle inequality for the distance function. In the piecewise approach it is immediate because the concatenation of a piecewise smooth curve from p to p' and from p' to p'' is again piecewise smooth. However the concatenation of smooth paths will in general not be a smooth path. Therefore the proof that this distance function is a distance function, Corollary 4.24 will have to be delayed until after we have proved

Corollary 4.23, the 'corner rounding' lemma. The proof that d^g is symmetric is trivial and the proof that it has positivity will be addressed in Corollary 4.18.

Remark 4.15. We see that the definition of the distance function implicitly assumes that there is a smooth path between every point. For manifolds, connected, path-connected, and smooth-path-connected are all equivalent. When we use the distance function, we assume that the manifold is connected.

We truly need to take the infimum, even in the euclidean case. Consider $U = \mathbb{R}^2 \setminus \{0\}$ and two points $p, -p$. The distance between them is $2\|p\|$. The unique path in the full plane that achieves this is the straight line, but this is not a path in the manifold, because it would pass through the removed origin. However, by taking paths that pass arbitrarily close to the origin, we see that the distance between these points is still $2\|p\|$

This leads to an important point: even for manifolds that are Riemannian immersed in euclidean space the distance function of Definition 4.13 is not the restriction of the euclidean distance function. Consider $U = \mathbb{R}^2 \setminus \overline{B(0,1)}$. Now the (limiting) shortest path between p and $-p$ skims the boundary of the unit disk An easy estimate shows that any such path is longer than $2\sqrt{1+\|p\|^2}$, which is strictly greater than $2\|p\|$. Therefore the distance function coming from the Riemannian metric and the restriction of the euclidean distance function to U are different. It is possible to show however that both generate the same topology.

Theorem 4.16. *The topology induced by d^g is equal to the topology of the M as a manifold.*

Proof. Choose a point p and a chart $U \subset \mathbb{R}^n$ that contains it. Within this chart, there is a convex compact neighbourhood K of p. In this chart we have the metric g_{ij} as a matrix of functions, but we can also consider the euclidean metric on this chart δ_{ij}. Consider all the vectors on K that are unit-length with respect to δ_{ij}. This is a compact set $K \times \mathbb{S}^{n-1}$. Because g is positive definite and continuous, it obtains a positive maximum C and minimum c on this set of vectors. For any vector $v \in TK$, write $v = \|v\|_{\text{euc}}\hat{v}$ with $\|\hat{v}\|_{\text{euc}} = 1$, then

$$\|v\|_{\text{euc}} c \leq \|v\|_g = \|v\|_{\text{euc}}\|\hat{v}\|_g \leq \|v\|_{\text{euc}} C.$$

If we apply this to paths, we obtain

$$cL^{\text{euc}}(\alpha) = c\int_a^b \|\alpha'\|_{\text{euc}}\, dt \leq L^g(\alpha) = \int_a^b \|\alpha'\|_g\, dt \leq C\int_a^b \|\alpha'\|_{\text{euc}}\, dt = CL^{\text{euc}}(\alpha). \tag{4.17}$$

Because K is convex, any two points can be joined with a straight-line, which achieves the euclidean distance between those points. Hence for any smooth path from p to q

$$c\, d^{\text{euc}}(p,q) \leq cL^{\text{euc}}(\alpha) \leq L^g(\alpha).$$

This proves that $c\, d^{\text{euc}}(p,q)$ is a lower bound for $d^g(p,q)$, and likewise $C\, d^{\text{euc}}(p,q)$ is an upper bound. Hence for small balls measured with d^g, there is a euclidean ball within it and containing it. Since balls generate the topology, this shows the two topology are equal. □

Corollary 4.18 (Positivity of Distance)**.** *$d^g(p,q) = 0$ if and only if $p = q$.*

Proof. One direction is easy: if $p = q$ take the constant path $\alpha(t) = p$. This has length zero.

Assume then that $d^g(p, q) = 0$. As discussed above we don't know if there is a minimum length path. Let α_n be a sequence of smooth paths from p to q whose lengths are decreasing to zero. Because the integral defining length is positive, any restriction of a curve decreases its length. Choose a chart containing U containing p and a compact set K as in the previous proof. There are two possibilities, either α_n never leaves K or it leaves K for the first time at $\alpha_n(t_n) \in \partial K$. If it never leaves K, set $t_n = b_n$ so that $\alpha_n(t_n) = q$. Hence we have a sequence of points

$$q_n = \alpha_n(t_n) = \begin{cases} q \\ \alpha_n(t_n) \in \partial K. \end{cases}$$

This is a sequence in the compact set K (if $q \notin K$, this is not a contradiction, this just means it will never occur in the sequence). Hence it has a convergent subsequence, whose limit q' is either q or a point in ∂K.

But now we know

$$d^{\text{euc}}(p, q') \leq c^{-1} d^g(p, q') \leq c^{-1} L^g(\alpha_n) \to 0.$$

Hence $p = q'$. Therefore either $p = q$ or $p \in \partial K$. But the second possibility is a contradiction, because p is in the interior of K by definition. □

Hopefully this is illustrative of the general strategy in this topic: We try to reduce the situation to a single chart, and then in the chart we can compare our metric g with the easily understood euclidean metric on the chart. With this in mind, we can now prove the corner rounding lemma. First we give an example in the plane.

Example 4.19 (Curve Blending)**.** Consider the piecewise smooth curve $\alpha : (-1, 1) \to \mathbb{R}^2$ given by

$$\alpha(t) = \begin{cases} (t, 0) & \text{for } t \leq 0 \\ (0, t) & \text{for } t > 0 \end{cases}$$

It has a corner at $t = 0$. We can think of this as the joining of the curve $\alpha_0(t) = (t, 0)$ and the curve $\alpha_1(t) = (0, t)$.

Let $\phi_\varepsilon(x)$ be a smooth step function that is 0 for $x < -\epsilon$ and 1 for $x > \varepsilon$. Then $\alpha_\varepsilon(t) := (1 - \phi_\varepsilon(t))\alpha_0(t) + \phi_\varepsilon(t)\alpha_1(t)$ is a smooth curve that agree with α for $t \notin (-\varepsilon, \varepsilon)$.

The curve in the above example had a right angle at its corner, but actually that was not particularly important. The key trick was to smoothly extend both sides of the curve past the corner. It wasn't even necessary that the two extensions intersected, the blending formula will still produce a smooth curve. So then how can we know that every smooth curve can be extended past the corner? We don't try to do it exactly, we allow ourselves to 'graft on' at a nearby point.

Example 4.20 (Curve Extension)**.** Take a smooth curve $\alpha : (a, 0) \to \mathbb{R}^n$. For any small $\varepsilon > 0$, let β_ε be the tangent line to α at $\alpha(-2\varepsilon)$. Then blending the curve with its tangent around the point of tangency produces

$$(1 - \phi_\varepsilon(t + 2\varepsilon))\alpha(t) + \phi_\varepsilon(t + 2\varepsilon)\beta_\varepsilon(t).$$

This is well-defined for all $t \in (a, \infty)$ It agrees with α for $t < -3\varepsilon$ and with β_ε for $t > -\varepsilon$.

Now we can use three blends to remove a corner from a piecewise smooth curve. In fact, performing this operation on smaller pieces of the curve ever closer to the corner results in curves that have almost the same length.

Lemma 4.21 (Corner Rounding, Euclidean)**.** *Let $\alpha : (a, b) \to \mathbb{R}^n$ be a piecewise smooth curve with a single corner at $t = 0$. Then for every $\varepsilon > 0$ there exists a smooth curve $\alpha_\varepsilon : (a, b) \to \mathbb{R}^n$ that agrees with α for $|t| > 2\varepsilon$. Moreover, $\lim_{\varepsilon \to 0} L(\alpha_\varepsilon) = L(\alpha)$.*

Proof. The previous two examples have shown us how to construct α_ε. First, on both sides of the corner, use Example 4.20 to extend $\alpha|_{(a,0)}$ and $\alpha|_{(0,b)}$ past 0. Then use Example 4.19 to blend the two together at 0. In more detail, for t in $(-3\varepsilon, -\varepsilon)$ we blended $\alpha|_{(a,0)}$ with its tangent β_- at $t = -2\varepsilon$. Similarly for t in $(\varepsilon, 3\varepsilon)$ we blended $\alpha|_{(0,b)}$ with its tangent β_+ at $t = 2\varepsilon$. And for $t \in (-\varepsilon, \varepsilon)$ we blended the two tangents together.

It only remains to show that this process hasn't increased the length too much. Because length is additive

$$\begin{aligned} L(\alpha_\varepsilon) - L(\alpha) &= \int_a^b \|\alpha_\varepsilon'\| - \|\alpha'\| \, dt = \int_{-3\varepsilon}^{3\varepsilon} \|\alpha_\varepsilon'\| - \|\alpha'\| \, dt \\ &= \left(\int_{-3\varepsilon}^{-\varepsilon} + \int_{-\varepsilon}^{\varepsilon} + \int_{\varepsilon}^{3\varepsilon} \right) \|\alpha_\varepsilon'\| - \|\alpha'\| \, dt \\ &\leq \left(\int_{-3\varepsilon}^{-\varepsilon} + \int_{-\varepsilon}^{\varepsilon} + \int_{\varepsilon}^{3\varepsilon} \right) \|\alpha_\varepsilon'\| \, dt. \end{aligned}$$

So we consider the three blends separately.

Let us calculate how much blending two curves α, β at $t = 0$ increases the length. For convenience, assume that the curves are parameterised by arc-length.

$$\begin{aligned} \gamma(t) &= (1 - \phi_\varepsilon(t))\alpha(t) + \phi_\varepsilon(t)\beta(t) \\ \gamma'(t) &= \phi'(\beta - \alpha) + (1 - \phi_\varepsilon)\alpha' + \phi_\varepsilon \beta' \\ \|\gamma'(t)\| &\leq \phi' \|\beta - \alpha\| + (1 - \phi_\varepsilon)1 + \phi_\varepsilon 1 \\ &= \phi' \|\beta - \alpha\| + 1, \end{aligned}$$

which gives

$$\int_{-\varepsilon}^{\varepsilon} \|\gamma'(t)\| \, dt \leq \int_{-\varepsilon}^{\varepsilon} \phi_\varepsilon' \|\beta - \alpha\| \, dt + 2\varepsilon \leq \sup_{t \in (-\varepsilon, \varepsilon)} \|\beta - \alpha\| + 2\varepsilon.$$

Because we have smooth curves, $\|\beta - \alpha\|^2$ is smooth and hence Lipschitz on any compact interval:

$$\left| \|\beta(t) - \alpha(t)\|^2 - \|\beta(0) - \alpha(0)\|^2 \right| < C|t|.$$

In total,

$$\int_{-\varepsilon}^{\varepsilon} \|\gamma'(t)\| \, dt \leq \sqrt{\|\beta(0) - \alpha(0)\|^2 + C_{\alpha,\beta}\varepsilon} + 2\varepsilon,$$

which tends to $\|\beta(0) - \alpha(0)\|$ as $\varepsilon \to 0$.

This shows that the blends of the curve with the two tangents tends to zero in the limit. But the two tangents may be skew to one another and not meet in $\mathbb{R}^n$. However, we can handle this with some triangle inequalities:

$$\begin{aligned}\|\beta_+(0) - \beta_-(0)\| &\leq \|\beta_+(0) - \beta_+(2\varepsilon)\| + \|\beta_+(2\varepsilon) - \beta_-(-2\varepsilon)\| + \|\beta_-(-2\varepsilon) - \beta_-(0)\| \\ &\leq 2\varepsilon + \|\alpha(2\varepsilon) - \alpha(-2\varepsilon)\| + 2\varepsilon,\end{aligned}$$

and since α is continuous, this goes to zero also. □

In a draft of this lemma, I tried to be explicit about the rate of convergence of the lengths but it made the proof even messier. I think the rate of convergence is $\mathrm{O}(\varepsilon)$, but this is only a guess.

Exercise 4.22. Determine the rate of convergence.

Another approach I considered for the corner rounding process was to use the curve shortening flow. Geometric flows are an important area of research. Most famously, there is the Ricci flow, which was used to solve the Poincare conjecture. In the curve shortening flow, one considers a family $\alpha_t(s)$ of arc-length parameterised curves such that

$$\frac{\partial \alpha_t}{\partial t} = \kappa_t(s) N_t(s).$$

for the curvature κ_t and normal N_t. If we think of t as a time parameter, it says that the curve is moving fastest where it is most curved, towards the center of the osculating circle. From our knowledge of curves and curvature, we know that this is equal to

$$\frac{\partial \alpha_t}{\partial t} = \frac{\partial^2 \alpha_t}{\partial s^2}.$$

We see that this is a heat equation. We know that the heat equation has very good regularity properties. Given a continuous initial condition, which in this case would be a curve α_0, the solution is smooth in t and analytic in s. Thus if we begin with a continuous curve, then we obtain a smoothing through this flow. I recommend Andrews et al to students who are intrigued by this.

Our 'elementary' corner rounding construction, though it seems as if it is particular to $\mathbb{R}^n$, can also be performed in a Riemannian manifold.

Corollary 4.23 (Corner Rounding, Riemannian)**.** *Let $\alpha : (a, b) \to M$ be a piecewise smooth curve in a Riemannian manifold M with a single corner at $t = 0$. Then for every $\varepsilon > 0$ there exists a smooth curve $\alpha_\varepsilon : (a, b) \to M$ that agrees with α for $|t| > 2\varepsilon$. Moreover, $\lim_{\varepsilon \to 0} L(\alpha_\varepsilon) = L(\alpha)$.*

Proof. The statement is local, in that we only need to change α in a neighbourhood of $\alpha(0)$. Choose a chart U containing $\alpha(0)$. Then by applying Lemma 4.21, we have a smooth curve α_ε. We know that its length converges to that of α in the euclidean metric of the chart, but not in the Riemannian metric g. As in the previous proof, it is enough to show that

$$L^g(\alpha_\varepsilon) - L^g(\alpha) = \int_{-3\varepsilon}^{3\varepsilon} \|\alpha_\varepsilon'\|_g - \|\alpha\|_g \, dt \leq \int_{-3\varepsilon}^{3\varepsilon} \|\alpha_\varepsilon'\|_g \, dt$$

tends to zero, but this follows immediately from the estimate (4.17). □

Corollary 4.24 (Triangle Inequality)**.** *On a Riemannian manifold, the distance function obeys the triangle inequality* $d^g(p,p'') \leq d^g(p,p') + d^g(p',p'')$.

Proof. As we already indicated, the proof comes down the class of curves used to define the distance function. We defined it with smooth curves. Consider any sequences of smooth paths α_ε from p to p' and β_ε from p' to p''. Assume that as $\varepsilon \to 0$ these approach the respective distances between the points.

Now concatenate the paths to get a piecewise smooth path with possibly a corner at p'. Apply the corner rounding procedure of Corollary 4.23 to obtain a sequence of smooth paths from p to p''. Let γ_ε be the ε-blending of α_ε and β_ε. In other words, we have taken the diagonal subsequence from the sequence of pairs of paths and the sequence of their blending. Therefore

$$d^g(p,p'') \leq \lim_{\varepsilon\to 0} L(\gamma_\varepsilon) = \lim_{\varepsilon\to 0} L(\alpha_\varepsilon) + L(\beta_\varepsilon) = d^g(p,p') + d^g(p',p''). \qquad \square$$

We close this section with a classic result: that geodesics are critical points of the length functional. The proof of this statement is reminiscent of the variational approach in Section 1.6 to show that minimal surfaces have vanishing mean curvature. In that situation, we considered graphs, so we could model a variation of the surface by adding another function. In the present case, we don't want to make a similar simplification and instead we will work directly with smooth families of paths.

Theorem 4.25 (First Variation of Length)**.** *Suppose that* $\alpha : (-\varepsilon,\varepsilon)\times[a,b] \to M$ *is a smooth family of paths. Let* $\alpha(s,a) = \alpha(s,0)$ *and* $\alpha(s,b) = \alpha(0,b)$ *so that all paths have the same endpoints. Thus this family of paths represents a variation of the path* $\alpha(0,t)$*. Without loss of generality, assume that* $t \mapsto \alpha(0,t)$ *is parameterised by arc-length. Then*

$$\frac{d}{ds} L(\alpha(s,\cdot))\Big|_{s=0} = -\int_a^b g(\partial_s\alpha, \nabla_{\partial_t\alpha}\partial_t\alpha)\, dt.$$

Therefore $\alpha(0,t)$ *is a geodesic if and only if it is a critical point of* L*.*

Proof. We compute

$$\begin{aligned}
\frac{d}{ds} L(\alpha(s,\cdot)) &= \int_a^b \frac{d}{ds}\sqrt{g(\partial_t\alpha,\partial_t\alpha)}\, dt \\
&= \int_a^b \frac{1}{2} g(\partial_t\alpha,\partial_t\alpha)^{-1/2} g(\partial_t\alpha, \nabla_{\partial_s\alpha}\partial_t\alpha)\, dt \\
&= \int_a^b \|\partial_t\alpha\|_g^{-1/2} g(\partial_t\alpha, \nabla_{\partial_s\alpha}\partial_t\alpha)\, dt.
\end{aligned}$$

When we set $s = 0$ into the above, $\|\partial_t\alpha\|_g^{-1/2} = 1$ because we assumed that $\alpha(0,t)$ was arc-length parameterised. Next we can use Lemma 3.44 to turn the covariant derivative in the $\partial_s\alpha$ into on in the $\partial_t\alpha$ direction. Additionally,

$$\frac{d}{dt} g(\partial_s\alpha,\partial_t\alpha) = g(\nabla_{\partial_t\alpha}\partial_s\alpha,\partial_t\alpha) + g(\partial_s\alpha, \nabla_{\partial_t\alpha}\partial_t\alpha).$$

Putting these together gives

$$\begin{aligned}\frac{d}{ds}L(\alpha(s,\cdot))\Big|_{s=0} &= \int_a^b g(\partial_t\alpha, \nabla_{\partial_t\alpha}\partial_s\alpha)\ dt \\ &= \int_a^b \frac{d}{dt}g(\partial_s\alpha, \partial_t\alpha) - g(\partial_s\alpha, \nabla_{\partial_t\alpha}\partial_t\alpha)\ dt \\ &= g(\partial_s\alpha, \partial_t\alpha)\Big|_a^b - \int_a^b \frac{d}{dt}g(\partial_s\alpha, \nabla_{\partial_t\alpha}\partial_t\alpha)\ dt.\end{aligned}$$

The first term vanishes because all the terms have the same endpoint, hence $\partial_s\alpha(s,a) = \partial_s\alpha(s,b) = 0$. □

That geodesics are critical points of length, not minima is significant.

Example 4.26. Consider the case of $\mathbb{S}^2$ and take points p, q that are not antipodes of one another. There is a unique great circle through these two points, and these points break it into two (arc-length parameterised) geodesic paths, one long and one short. The shorter geodesic is the minimum (we will prove this in the next section) but the longer geodesic is a saddle point in the space of smooth paths between these points. We can see that it is a saddle point by imagining variations. If most of the path is fixed, and we just add a variation in one small area then the geodesic will have the lowest length of this family of paths. On the other hand, consider all the planes that contain p, q. You can do this by rotating a plane on the line through p, q. If you take the family of paths that are the intersection of these planes and the sphere, then the long geodesic is the longest and the short geodesic is the shortest path between p, q in this family.

4.4 Exponential Maps

Although we have given a reasonable definition of the distance on a Riemannian manifold, it is often very difficult in practice to understand this function. We have not given an example of the distance between two points yet because the prospect of searching for the infimum of length over every possible path is daunting. Even in a well-understood space such as the euclidean plane, where would you even begin? While at first glance it seems as if Theorem 4.25 simplifies the search to geodesics, this is only a complete answer if you already know that there exists a *length-minimising* path between the two points. Again, the example of a punctured plane shows that a length-minimising path may not exist. The example of a punctured sphere and two points either side of the puncture, such that the shorter geodesic is blocked, shows that even though the two points are connected by a geodesic (the longer geodesic), its length is not the distance.

However, what we will show in this chapter is every point has a special ball around it: every point in that ball has a unique geodesic to the center and the length of this geodesic is the distance to the center. The key gadget for this construction is the *exponential map*, which can be summarised as 'follow the geodesic out from the center'. The name is due to a relationship with the exponential map in Lie group theory and actual exponentials will not be relevant to us here.

Definition 4.27. *For any $p \in M$ let the* exponential map $\exp_p : T_pM \rightharpoonup M$ *be the function defined by*

$$\exp_p(v) = \gamma_{p,v}(1),$$

where $\gamma_{p,v}$ is the unique maximal geodesic with $\gamma_{p,v}(0) = 0$ and $\gamma'_{p,v}(0) = v$. Because the solutions of ODEs depend smoothly on their initial conditions, $\exp_p$ is a smooth function of v.

Recall, due to the Rescaling Lemma 4.2 geodesics that only differ by speed have the same image. In the exponential map we have removed this duplication by only considering $t = 1$. This is the only advantage of defining the exponential map over using geodesics directly. In fact the two are equivalent, since from the Rescaling Lemma we have that

$$\gamma_{p,v}(t) = \gamma_{p,tv}(1) = \exp_p(tv).$$

The exponential map is a partially defined function; its value only exists if the corresponding geodesic exists at time $t = 1$. Geodesics are defined by an ODE with short-time existence, but we know nothing about their long-time existence. It is natural therefore to ask about $\operatorname{dom}\exp_p \subset T_pM$. Because of the constant curve $\alpha(t) = p$ we know that $0 \in \operatorname{dom}\exp_p$. Likewise, short-time existence tells us that $\operatorname{dom}\exp_p$ contains a neighbourhood of 0. The final thing we can say is that $\operatorname{dom}\exp_p$ is star-shaped around $0 \in T_pM$: For clarity, write $t = c$ in the above equation: $\exp_p(cv) = \gamma_{p,v}(c)$. If $\gamma_{p,v}(1)$ exists then $\gamma_{p,v}(c)$ exists for $c < 1$ because $\gamma_{p,v}$ is maximal. This shows that if $v \in \operatorname{dom}\exp_p$ then $cv \in \operatorname{dom}\exp_p$ for $0 \leq c \leq 1$.

The next thing we have to understand is tricky because it breaks the usual hierarchy of concepts. We need to think about the tangent map $T_v \exp_p : T_v(\operatorname{dom}\exp_p) \to T_{\exp_p(v)}M$. In particular, for $v = 0$ it is a map between T_pM and itself. Recall the definition of a manifold in Chapter 2. A chart U is an open subset of $\mathbb{R}^n$ and the tangent space is $T_pM = \mathbb{R}^n$ (with an equivalence relation between these $\mathbb{R}^n$ for different charts). Likewise we can think of $\operatorname{dom}\exp_p \subset T_pM$ as

an open subset of euclidean space and $T_v(\operatorname{dom}\exp_p)$ as $\mathbb{R}^n$. Given $w \in T_pM$ we think of it in $T_v(\operatorname{dom}\exp_p)$ as the curve $v + tw$. For any $w \in T_pM$ we have

$$T_0 \exp_p(w) = \left.\frac{d}{dt}\exp_p(0+tw)\right|_{t=0} = \left.\frac{d}{dt}\gamma_{p,w}(t)\right|_{t=0} = w.$$

Hence $T_0\exp_p = \mathrm{id}_{T_pM}$.

Because $T_0\exp_p$ is an isomorphism, the inverse function theorem says that there is a neighbourhood $V \ni 0$ such that $\exp_p|_V$ is an diffeomorphism onto its image. We can further restrict V so that the image is entirely within the chart U[4]. Therefore we have a transition function $\exp_p|_V : V \to U$. Because $\operatorname{img}\exp_p$ is entirely within U, it will obey the cocycle conditions with every other transition function of M. Because it is defined on all of V adding this chart to the atlas does not create any new points of M. In the literature this construction is called *normal coordinates* at p. We will call this chart a *normal chart* at p.

A normal chart has two metrics on it. Because it is a chart of M, of course it has the metric g. But V is also a subset of T_pM and T_pM is an inner-product space using the inner product $g|_p$. Therefore V also has the metric that just uses $\tilde{g} = g|_p$ at every point. Clearly the two metrics are equal at p. The coefficients $\tilde{g}_{ij}$ of this second metric are constants, and the second metric can be thought of as a euclidean metric. The natural question is therefore how g and $\tilde{g}$ compare.

A normal chart V is centered on p, in the sense that $[0_V] = p \in M$. The most useful property of a normal chart is that rays $t \mapsto tv$, which are geodesics of $\tilde{g}$, are geodesics of g. This is because $\tilde{g}_{ij}$ is constant, so its geodesics are straight lines in V. On the other hand, if we view these rays in the chart U, we have $t \mapsto \exp_p(tv) = \gamma_{p,v}(t)$, which are by definition geodesics of g. For this reason we call these rays *radial geodesics*, without needing to specify which metric. Similarly we define *geodesic balls* and *geodesic spheres* centered at p to be the sets

$$\tilde{B}_r = \{v \in T_pM \mid \|v\|_{\tilde{g}} < r\}, \qquad \partial\tilde{B}_r = \{v \in T_pM \mid \|v\|_{\tilde{g}} = r\}.$$

By restricting V we may assume that it is a geodesic ball. Of course rays and sphere are orthogonal with respect to $\tilde{g}$ but remarkably:

Lemma 4.28 (Gauss' Lemma)**.** *Radial geodesics are orthogonal to geodesic spheres with respect to the metric g.*

Proof. Choose any smooth function $\omega : (-\varepsilon, \varepsilon) \to T_pM$ with $\|\omega\|_{\tilde{g}} = 1$ and consider the family of curves in V

$$\alpha(s,t) = t\omega(s).$$

This family has the property that every main curves $\alpha(s,\cdot)$ is a radial geodesic and that every transverse curves $\alpha(\cdot,t)$ lies in a geodesic sphere. This means that $\partial_s\alpha(0,t)$ is a tangent vector to the geodesic sphere $\partial\tilde{B}_t$. Conversely, given any tangent vector to a geodesic sphere arises in this way.

The lemma comes down to showing that $\partial_t\alpha(0,t)$ and $\partial_s\alpha(0,t)$ are orthogonal for all t with respect to g. Because $\partial_s\alpha(0,0) = 0\omega'(0) = 0$ it is enough to prove that $g(\partial_t\alpha(0,t), \partial_s\alpha(0,t))$ is

[4]It is not necessary to do this step, but it makes things simpler.

constant. We compute, for ∇ the Levi-Civita connection of g,

$$\begin{aligned}\frac{\partial}{\partial t} g(\partial_t\alpha(0,t), \partial_s\alpha(0,t)) &= g(\nabla_{\partial_t\alpha}\partial_t\alpha(0,t), \partial_s\alpha(0,t)) + g(\partial_t\alpha(0,t), \nabla_{\partial_t\alpha}\partial_s\alpha(0,t)) \\ &= 0 + g(\partial_t\alpha(0,t), \nabla_{\partial_s\alpha}\partial_t\alpha(0,t)) \\ &= \frac{\partial}{\partial s}\frac{1}{2} g(\partial_t\alpha(0,t), \partial_t\alpha(0,t)) = 0.\end{aligned}$$

The facts we have used here are that a geodesic parallel transports its tangent vector (Definition 4.1), a geodesic is constant speed (Lemma 4.6), and symmetry of covariant derivatives for families of curves (Lemma 3.44). □

Example 4.29. How does this apply to the hyperbolic plane? The hyperbolic plane is covered by a single chart $U = \mathbb{H}^2$. Consider the set of geodesics through any point p. We know that they are defined for all time, positive and negative. This tells us that $\exp_p$ is defined on the whole tangent space. We also know that for any point q there is a geodesic from p to q. This tells us that $\exp_p$ is surjective onto U. Finally, these geodesics only intersect at p. This means that $\exp_p$ is diffeomorphism from T_pM to U, and the normal chart at p covers all of $\mathbb{H}^2$.

Let us be explicit for the point $p = (0,1)$. The geodesics though this point have the form $\gamma(t) = (0, e^t)$ or

$$\gamma(t) = (a^2+1)\left(\frac{a}{a^2+e^{2t}} - \frac{a}{a^2+1}, \frac{e^t}{a^2+e^{2t}}\right) = \left(\frac{a}{a^2+e^{2t}}(1-e^{2t}), \frac{a^2+1}{a^2+e^{2t}}e^t\right).$$

We should try to combine these into a single formula. In the limit $a \to \pm\infty$ we have $\gamma(t) \to (0, e^t)$. Therefore the problem is that a is not parameterised correctly in some sense. Let $a = \cot\theta/2$. Then

$$\begin{aligned}\sin^2\frac{\theta}{2}(a^2+e^{2t}) = \sin^2\frac{\theta}{2}(\csc^2\frac{\theta}{2} - 1 + e^{2t}) &= 1 + \frac{1}{2}(1-\cos\theta)(-1+e^{2t}) \\ &= \frac{1}{2}[(1+\cos\theta) + (1-\cos\theta)e^{2t}]\end{aligned}$$

and

$$\begin{aligned}\gamma(t) &= \frac{2\sin^2\frac{\theta}{2}}{(1+\cos\theta)+(1-\cos\theta)e^{2t}}\left(\frac{\cos\frac{\theta}{2}}{\sin\frac{\theta}{2}}(1-e^{2t}), \csc^2\frac{\theta}{2}e^t\right) \\ &= \frac{1}{(1+\cos\theta)+(1-\cos\theta)e^{2t}}\left(\sin\theta(1-e^{2t}), 2e^t\right).\end{aligned}$$

Now we see that we have a nice parameterisation $\gamma_\theta(t)$ of the geodesics through $(0,1)$, with θ giving the angle of the tangent with ∂_2. For example $\theta = 0$ gives $\gamma_0(t) = (0, e^t)$ and $\theta = \pi$ gives $\gamma_\pi(t) = (0, e^{-t})$. More generally we see that

$$\gamma_\theta(-t) = \frac{1}{(1+\cos\theta)e^{2t} + (1-\cos\theta)}\left(\sin\theta(e^{2t}-1), 2e^t\right) = \gamma_{\theta+\pi}(t).$$

In fact, (t,θ) are polar coordinates for the normal chart at $(0,1)$.

This gives us a formula for geodesic spheres. In normal coordinates of course, geodesic spheres around p are just $t = r$. In (x, y) coordinates, they are

$$\alpha(\theta) = \frac{1}{(1+\cos\theta)+(1-\cos\theta)R^2}\left(\sin\theta(1-R^2), 2R\right),$$

for $R = e^r$. These are euclidean circles with centers at $(0, \cosh R)$ and radii $\sinh R$.

Example 4.30. We can apply the same analysis to the sphere. Consider the set of geodesics through any point p. We know that they are defined for all time, positive and negative. This tells us that $\exp_p$ is defined on the whole tangent space. We also know that for any point q there is a geodesic from p to q. This tells us that $\exp_p$ is surjective onto U. However, this time all geodesics from p intersect at the antipode of p. Therefore $\exp_p$ is not injective. To construct a normal chart we therefore have to restrict to a subset of T_pM.

For the south pole, to choose a definite point, we know the geodesics are the lines of longitude. Therefore the normal chart at this point has to have lines of longitude as rays. The normal chart is in fact just a rescaling of U_N (stereographic projection) so that the rays are unit speed geodesics.

Theorem 4.31. *For any $p \in M$ consider any other point q that lies in a geodesic ball of p. Let $\alpha : [a, b] \to M$ be any smooth path from p to q, not necessarily lying in the geodesic ball. Then the radial geodesic from p to q is the unique length-minimiser from p to q.*

Proof. We again use a normal chart $V = \tilde{B}_{r'}$ at p. The radial geodesic has the form $\gamma(t) = t\omega$ for some $\|\omega\|_{\tilde{g}} = 1$. The point $q = r\omega$ lies on the geodesic sphere $\partial\tilde{B}_r \subset \tilde{B}_{r'}$. Since v is tangent vector of γ at p and geodesics are constant speed

$$\|\gamma'\|_g = \|\gamma'(0)\|_g = \|\gamma'(0)\|_{\tilde{g}} = \|\omega\|_{\tilde{g}} = 1.$$

Therefore

$$L(\gamma) = \int_0^r \|\gamma'\|_g \, dt = \int_0^r 1 \, dt = r.$$

Now suppose first that α stays entirely within the geodesic ball $\tilde{B}_{r'}$. This means we can write $\alpha(t) = \rho(t)\omega(t)$, where $\rho : [a, b] \to \mathbb{R}_{\geq 0}$ and $\omega : [a, b] \to T_pM$ with $\|\omega(t)\|_{\tilde{g}} = 1$. So notice that $\rho\omega'$ is tangent to the geodesic sphere $\partial\tilde{B}_\rho$. Also for any t we can consider the radial geodesic $s \mapsto s\omega(t)$. It follows that $g(\omega, \omega) = 1$ at any point of V. Then with the help of Gauss' Lemma 4.28

$$\begin{aligned}
\|\alpha'(t)\|_g^2 &= g|_{\alpha(t)}(\rho'\omega + \rho\omega', \rho'\omega + \rho\omega') \\
&= (\rho')^2 g|_{\alpha(t)}(\omega, \omega) + 2\rho\rho' g|_{\alpha(t)}(\omega, \omega') + \rho^2 g|_{\alpha(t)}(\omega', \omega') \\
&= (\rho')^2 + 0 + \rho^2 g|_{\alpha(t)}(\omega', \omega') \\
&\geq (\rho')^2.
\end{aligned}$$

Therefore

$$L(\alpha) = \int_a^b \|\alpha'(t)\|_g \, dt \geq \int_a^b \rho'(t) \, dt = \rho(b) - \rho(a) = r - 0 = L(\gamma).$$

Finally, if α does not stay entirely within the geodesic ball $\tilde{B}_{r'}$, there there must be some first time $t = \tilde{b}$ that it crosses $\partial\tilde{B}_r$. Then

$$L(\alpha) \geq L(\alpha|_{[a,\tilde{b}]}) \geq L(\gamma).$$

Thus the radial geodesic is a length-minimiser from p to q.

For the converse, observe that it is only possible to have equality if α lies entirely within the geodesic ball and $g|_{\alpha(t)}(\omega', \omega') = 0$ for all t. This implies that $\omega' = 0$ and that α is radial. □

Remark 4.32. An immediate consequence of this is that geodesic balls and metric balls are the same sets, and the r of the geodesic ball really is its radius.

Example 4.33. From Example 4.29 we know that for any point the normal chart covers the entire of the hyperbolic plane. Therefore the distance between any two points is given by the length of the geodesic path between them. This was calculated for vertical geodesics in Example 4.14. We can develop this into a general distance formula.

Consider two points $p = (x_1, y_1), q = (x_2, y_2) \in \mathbb{H}^2$. We will use isometries to bring them both to the y-axis. Firstly we translate so that $p' = (0, y_1), q' = (w, y_2)$ for $w = x_2 - x_1$. Next, we construct an isometry using the complex number form. Suppose that f transforms p' to $(0, 1) = \iota$ and q' to $(0, r) = \iota r$. Then the inverse transform f^{-1} obeys

$$\frac{a\iota + b}{c\iota + d} = p' = \iota y_1, \qquad \frac{a\iota r + b}{c\iota r + d} = q' = w + \iota y_2.$$

From the left equation, we see that $a = y_1 d$ and $b = -y_1 c$. Continuing with the right equation, we have

$$\begin{aligned}
\iota y_1 r d - y_1 c &= wd - y_2 rc + \iota(y_2 d + wrc), \\
\text{Re}: \quad 0 &= (y_1 - y_2 r)c + wd \\
\text{Im}: \quad 0 &= wrc + (y_2 - y_1 r)d, \\
\Rightarrow \quad 0 &= (y_2 - y_1 r)(y_1 - y_2 r) - w^2 r \\
&= y_1 y_2 - (y_1^2 + y_2^2 + w^2)r + y_1 y_2 r^2 \\
b &:= \frac{y_1^2 + y_2^2 + w^2}{2y_1 y_2} \\
r &= b \pm \sqrt{b^2 - 1}.
\end{aligned}$$

Thus the distance between p and q is given by $\ln\left|\frac{r}{1}\right| = \ln(b + \sqrt{b^2 - 1})$. In the case that $x_1 = x_2$ we see that $b^2 - 1 = (y_2^2 - y_1^2)^2/(2y_1 y_2)^2$ so that the distance formula reduces to

$$\ln \frac{y_1^2 + y_2^2 + y_2^2 - y_1^2}{2y_1 y_2} = \ln \frac{y_2}{y_1}$$

as expected from Exercise 4.14.

One can continue to torture the distance formula until it yields to geometric interpretation:

$$\begin{aligned}
&4y_1y_2(b+\sqrt{b^2-1}) \\
&= 2y_1^2+2y_2^2+2w^2+2\sqrt{(y_1^2+y_2^2+w^2)^2-4y_1^2y_2^2} \\
&= 2y_1^2+2y_2^2+2w^2+2\sqrt{(y_1^2-2y_1y_2+y_2^2+w^2)(y_1^2+2y_1y_2+y_2^2+w^2)} \\
&= (y_1-y_2)^2+(y_1+y_2)^2+2w^2+2\sqrt{((y_1-y_2)^2+w^2)((y_1+y_2)^2+w^2)} \\
&= \left(\sqrt{(y_1-y_2)^2+w^2}+\sqrt{(y_1+y_2)^2+w^2}\right)^2.
\end{aligned}$$

We have a the root of a sum of squares, so this must be some euclidean distance in the hyperbolic plane. Indeed, the left square root is the distance between p and q. For the other square root, if we take p, q, and the semicircle they lie on, and reflect those in the x-axis to make a full circle then the second square root is the distance between p and the reflection of q.

There is one more property of normal charts that we will need in the next chapter, namely that the metric in a normal chart has a nice form at p. For any chart, because $g|_p$ is non-degenerate we can always find an orthonormal basis of T_pM. We can then make a linear change of coordinates such that this basis is a coordinate vector basis near p. Doing this will diagonalise the metric, so $g_{ij}(p) = \delta_{ij}$. But typically g_{ij} are non-constant, so this property only holds at p and $\partial_k g_{ij}(p) \neq 0$. A special feature of normal charts is that the partial derivatives of the metric coefficients are zero, so that this procedure simplifies g_{ij} at p up to second order.

Lemma 4.34 (Metric in Normal Chart)**.** *In a normal chart at $p \in M$, all first partial derivatives of the metric coefficients vanish at p, i.e. $\partial_k g_{ij}(p) = 0$.*

Proof. In a normal chart at p, the point p is the origin and we know that the radial lines are geodesics. Let $\gamma(t) = tv$ be the geodesic in the v direction. If we put this into the geodesic equation,

$$0 + \Gamma^k_{ij}(tv)v^iv^j = 0.$$

In particular, if we put $t = 0$ into this equation and consider all possible v, this equation can only hold if $\Gamma^k_{ij}(0) = 0$ for all indices. From the formula (3.49) for the Christoffel coefficients of the Levi-Civita connection, this in turn is only possible if

$$\partial_i g_{kj}(0) + \partial_j g_{ik}(0) - \partial_k g_{ij}(0) = 0.$$

Permute the indices

$$\partial_j g_{ik}(0) + \partial_k g_{ji}(0) - \partial_i g_{jk}(0) = 0,$$

and add the two expressions together to obtain

$$2\partial_j g_{ik}(0) = 0.$$

□

Example 4.35. Consider $p = (0, 1) \in \mathbb{H}^2$. In the usual coordinates, $g_{ij} = y^{-2}\delta_{ij}$ so $g_{ij}(p) = \delta_{ij}$. But $\partial_2 g_{ij} = -2y^{-3}\delta_{ij}$ so $\partial_2 g_{ij}(p) \neq 0$.

Recall Example 4.29. In that exercise we constructed polar coordinates for a normal chart at p. Namely (t, θ) were related to the usual coordinates by

$$\begin{aligned}(x^1, x^2) = (x, y) &= \frac{1}{(1+\cos\theta) + (1-\cos\theta)e^{2t}} \left(\sin\theta(1-e^{2t}), 2e^t\right) \\ &= \frac{1}{\cosh t - \cos\theta \sinh t} \left(-\sin\theta \sinh t, 1\right).\end{aligned}$$

Let (u^1, u^2) be (cartesian) normal coordinates at p with $(u^1, u^2) = (t\cos\theta, t\sin\theta)$. Then we can calculate the metric in these normal coordinates using that change of chart formula from Section 3.1:

$$\tilde{g}_{ij} = \frac{\partial x^k}{\partial u^i}\frac{\partial x^l}{\partial u^j} g_{kl} = \frac{\partial x^k}{\partial u^i}\frac{\partial x^l}{\partial u^j} y^{-2}\delta_{kl} = \left(\frac{\partial x^1}{\partial u^i}\frac{\partial x^1}{\partial u^j} + \frac{\partial x^2}{\partial u^i}\frac{\partial x^2}{\partial u^j}\right) y^{-2}$$

What we want to see is that $\tilde{g}_{ij}(p) = \delta_{ij}$ and $\partial_k \tilde{g}_{ij}(p) = 0$. The direct approach, while elementary, is ugly. To make our point it is sufficient to give Taylor series for the coefficients of the metric up to first order, which requires in turn the Taylor series of the change of coordinates up to second order. The key function to understand is the denominator

$$d := \cosh t - \cos\theta \sinh t = \cosh t - t\cos\theta \frac{\sinh t}{t}.$$

Both $\cosh t$ and $\frac{\sinh t}{t}$ are even analytic functions of t. Therefore inserting $t = \sqrt{(u^1)^2 + (u^2)^2}$ gives

$$\begin{aligned} d &= \left(1 + \frac{1}{2!}t^2 + \dots\right) - u^1\left(1 + \frac{1}{3!}t^2 + \dots\right) \\ &= \left(1 + \frac{1}{2!}((u^1)^2 + (u^2)^2) + \dots\right) - u^1\left(1 + \frac{1}{3!}((u^1)^2 + (u^2)^2) + \dots\right) \\ &= 1 - u^1 + \frac{1}{2}(u^1)^2 + \frac{1}{2}(u^2)^2 + \dots, \end{aligned}$$

an analytic function of the normal coordinates. As $y = d^{-1}$, we can easily write down the Taylor series for y^{-2}

$$y^{-2} = 1 - 2u^1 + \dots.$$

For y itself, clearly $y(p) = 1$ and

$$\begin{aligned}
\frac{\partial y}{\partial u^1} &= -\frac{1}{d^2}\frac{\partial d}{\partial u^1} & \frac{\partial y}{\partial u^1}(p) &= 1\\
\frac{\partial y}{\partial u^2} &= -\frac{1}{d^2}\frac{\partial d}{\partial u^2} & \frac{\partial y}{\partial u^2}(p) &= 0\\
\frac{\partial^2 y}{\partial u^1 \partial u^1} &= \frac{2}{d^3}\left(\frac{\partial d}{\partial u^1}\right)^2 - \frac{1}{d^2}\frac{\partial^2 d}{\partial u^1 \partial u^1} & \frac{\partial^2 y}{\partial u^1 \partial u^1}(p) &= 1\\
\frac{\partial^2 y}{\partial u^1 \partial u^2} &= \frac{2}{d^3}\frac{\partial d}{\partial u^1}\frac{\partial d}{\partial u^2} - \frac{1}{d^2}\frac{\partial^2 d}{\partial u^1 \partial u^2} & \frac{\partial^2 y}{\partial u^1 \partial u^2}(p) &= 0\\
\frac{\partial^2 y}{\partial u^2 \partial u^2} &= \frac{2}{d^3}\left(\frac{\partial d}{\partial u^2}\right)^2 - \frac{1}{d^2}\frac{\partial^2 d}{\partial u^2 \partial u^2} & \frac{\partial^2 y}{\partial u^2 \partial u^2}(p) &= -1.
\end{aligned}$$

So the Taylor series of y up to second order is

$$y = 1 + u^1 + \frac{1}{2}(u^1)^2 - \frac{1}{2}(u^2)^2.$$

Next $x = -d^{-1}\sin\theta \sinh t = -yu^2\,\frac{\sinh t}{t}$ so up to second order

$$x = -\left(1 + u^1 + \dots\right) u^2 \left(1 + \dots\right) = -u^2 - u^1 u^2.$$

Finally we have up to first order

$$\begin{aligned}
\tilde{g}_{1\,1} &= \left(\left(\frac{\partial x}{\partial u^1}\right)^2 + \left(\frac{\partial y}{\partial u^1}\right)^2\right) y^{-2} = \left(\left(-u^2\right)^2 + \left(1 + u^1\right)^2\right)\left(1 - 2u^1\right)\\
&= (1 + 2u^1)(1 - 2u^1) = 1\\
\tilde{g}_{1\,2} &= \left(\frac{\partial x}{\partial u^1}\frac{\partial x}{\partial u^2} + \frac{\partial y}{\partial u^1}\frac{\partial y}{\partial u^2}\right) y^{-2} = \left((-u^2)(-1 - u^1) + (1 + u^1)(-u^2)\right)\left(1 - 2u^1\right)\\
&= \left(u^2 - u^2\right)\left(1 - 2u^1\right) = 0\\
\tilde{g}_{2\,2} &= \left(\left(\frac{\partial x}{\partial u^2}\right)^2 + \left(\frac{\partial y}{\partial u^2}\right)^2\right) y^{-2} = \left((-1 - u^1)^2 + (-u^2)^2\right)\left(1 - 2u^1\right)\\
&= \left(1 + 2u^1\right)\left(1 - 2u^1\right) = 1.
\end{aligned}$$

We will finish this chapter by indicating various directions in which one could continue. One can investigate further with Riemannian manifolds as metric spaces. We have encountered several times the example of the punctured plane and how it 'blocks' geodesics. The Hopf-Rinow theorem states that a connected Riemannian manifold is complete as a metric space if and only if every geodesic exists for all time. One can also develop the theory of the exponential map further. We have mentioned its connection to the exponential map in Lie group theory, so this could be expounded. We can also ask at every p what is the largest geodesic ball on which $\exp_p$ is injective, called the *injectivity radius*. Relatedly, we have the *cut locus* at p which asks when geodesics from p stop being the length-minimisers. We can also consider families of geodesics, using the *Jacobi field*, of which the radial geodesics are one example. Another example would be geodesics beginning on some hypersurface. A typical question is to ask at what rate these geodesics are moving apart.

Naturally one can look for special manifolds. We found the geodesics of the hyperbolic plane using isometries. As mentioned, spaces whose isometries are transitive are called homogeneous, and one whose isometries are transitive on the unit sphere of T_pM are called isotropic at p. Spaces such that for every point there is an isometry that acts as $-\mathrm{id}$ on T_pM are called symmetric spaces, and there is a complete classification.

These are all relatively 'pure' Riemannian geometry questions, in that they try to understand the intrinsic structure of a Riemannian manifold. But we can also take Riemannian manifolds as the setting to investigate all types of geometric problems. Geometric flows are one example. The final direction we will mention however is *harmonic maps*. Both minimal surfaces and geodesics are extremal for a functional, surface area and length respectively. If we model both of these problems as embedding stretched rubber objects and seeing the 'minimal tension' configuration, then this motivates the definition of a harmonic map. The name is due to the defining equation being a generalisation of the Laplacian to the context of Riemannian manifolds (compare the the Laplacian in Equation (1.37)). My PhD work was on harmonic maps from a torus into $\mathbb{S}^3$.

Chapter 5

Curvature

Curvature in Riemannian geometry can seem a little hidden, but we have already encountered its effects. We saw for instance that parallel transport around a loop on the sphere changes a vector. We also saw for the hyperbolic plane that triangles have angles that sum to less than pi. Both of these are due to the intrinsic curvature of the space. Our own universe has curvature: in general relativity it is curvature that causes gravity. We commonly interpret an asteroid being deflected as it passes a planet as it being pulled from its straight line path by a force; in truth it is travelling on a geodesic and it is space itself that is bent.

5.1 Symmetries and Identities

We first motivate curvature by locally comparing a Riemannian manifold M to euclidean space $\mathbb{R}^n$. Later we will connect it to the more geometric picture presented in Chapter 1. An important feature of euclidean space is that it has (a basis of) parallel vector fields with respect to the Levi-Civita connection. These are vector fields that are parallel along every curve. Because the Levi-Civita connection is uniquely determined by the metric, the property of having a parallel vector field must be a local isometry invariant, i.e. if M at p has a neighbourhood that is isometric to a neighbourhood of $\mathbb{R}^n$ then it has a parallel vector field.

The obvious way to construct a parallel vector field is to begin with a vector $Z|_p \in T_pM$ and parallel transport it around. Choose two coordinate direction ∂_1, ∂_2. We parallel transport $Z|_p$ along the x^1-axis and then from every point in the x^2 direction. Is the vector field so constructed parallel in the x^1 direction? By construction $\nabla_{\partial_1} Z = 0$ on the x^1-axis, so it is sufficient to have $\nabla_{\partial_2}\nabla_{\partial_1} Z = 0$.

An alternative way to ask this question is to construct a second vector field $\tilde{Z}$ with $\tilde{Z}|_p = Z_p$ by first parallel transporting $\tilde{Z}|_p$ along the x^2-axis and then from every point in the x^1 direction. By definition $\tilde{Z}$ and Z agree on the x^1- and x^2-axes. But do they agree at other points? Consider

a point $q = p + (h_1, h_2)$ close to p. Then using an approximation

$$\begin{aligned} Z|_{p+(h_1,0)} &\approx Z|_p + h_1(\nabla_{\partial_1} Z)|_p \\ Z|_q &\approx Z|_{p+(h_1,0)} + h_2(\nabla_{\partial_2} Z)|_{p+(h_1,0)} \\ &= \big[Z + h_1(\nabla_{\partial_1} Z)\big]|_p + h_2\big[\nabla_{\partial_2} Z + h_1(\nabla_{\partial_2}\nabla_{\partial_1} Z)\big]|_{p+(h_1,0)} \\ &= Z|_p + h_1 0 + h_2 0 + h_1 h_2(\nabla_{\partial_2}\nabla_{\partial_1} Z)|_{p+(h_1,0)}. \end{aligned}$$

Likewise

$$\tilde{Z}|_q = \tilde{Z}|_p + h_1 0 + h_2 0 + h_1 h_2(\nabla_{\partial_1}\nabla_{\partial_2} \tilde{Z})|_{p+(0,h_2)}.$$

For h_1, h_2 small and since $\tilde{Z}|_p = Z|_p$ we see that $\tilde{Z}|_q = Z|_q$ are equal if and only if $\nabla_{\partial_2}\nabla_{\partial_1} Z = \nabla_{\partial_1}\nabla_{\partial_2} Z$. Thus the lack of a parallel vector field can be measured by the difference $\nabla_{\partial_1}\nabla_{\partial_2} Z - \nabla_{\partial_2}\nabla_{\partial_1} Z$. We can move away from coordinates by using the properties of the connection to give an equivalent statement for arbitrary vector fields. For $X = X^i\partial_i, Y = Y^j\partial_j$

$$\begin{aligned} \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z &= X^i\nabla_{\partial_i}(Y^j\nabla_{\partial_j} Z) - Y^j\nabla_{\partial_j}(X^i\nabla_{\partial_i} Z) \\ &= X^i\frac{\partial Y^j}{\partial x^i}\nabla_{\partial_j} Z + X^iY^j\nabla_{\partial_i}\nabla_{\partial_j} Z - Y^j\frac{\partial X^i}{\partial x^j}\nabla_{\partial_i} Z - X^iY^j\nabla_{\partial_j}\nabla_{\partial_i} Z \\ &= X^iY^j\big(\nabla_{\partial_i}\nabla_{\partial_j} Z - \nabla_{\partial_j}\nabla_{\partial_i} Z\big) + \nabla_{X^i(\partial_i Y^j)\partial_j} Z - \nabla_{Y^j(\partial_j X^i)\partial_i} Z \\ &= X^iY^j\big(\nabla_{\partial_i}\nabla_{\partial_j} Z - \nabla_{\partial_j}\nabla_{\partial_i} Z\big) + \nabla_{X^i(\partial_i Y^j)\partial_j - Y^j(\partial_j X^i)\partial_i} Z \\ &= X^iY^j\big(\nabla_{\partial_i}\nabla_{\partial_j} Z - \nabla_{\partial_j}\nabla_{\partial_i} Z\big) + \nabla_{[X,Y]} Z. \end{aligned}$$

Just as for torsion, we see that this 'commutator' of vector fields has a part that is due the commutator of the vector fields themselves but also a part that is 'built in' to all such commutators.

Definition 5.1. *The* Riemannian curvature tensor R *is a vector valued function*

$$R(X,Y)Z = \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]} Z.$$

We see immediately from the calculation above that R only depends on the pointwise values of X and Y. We can continue to get an formula for the curvature tensor in terms of the Christoffel coefficients.

$$\begin{aligned} R(\partial_i,\partial_j)Z &= \nabla_{\partial_i}\nabla_{\partial_j}(Z^k\partial_k) - \nabla_{\partial_j}\nabla_{\partial_i}(Z^k\partial_k) \\ &= \nabla_{\partial_i}\big(\partial_j Z^k\partial_k + Z^k\nabla_{\partial_j}\partial_k\big) - \nabla_{\partial_j}\big(\partial_i Z^k\partial_k + Z^k\nabla_{\partial_i}\partial_k\big) \\ &= \nabla_{\partial_i}\big(\partial_j Z^k\partial_k + Z^k\Gamma^l_{jk}\partial_l\big) - \nabla_{\partial_j}\big(\partial_i Z^k\partial_k + Z^k\Gamma^l_{ik}\partial_l\big) \\ &= \big(\partial_i\partial_j Z^k\partial_k + \partial_j Z^k\Gamma^l_{ik}\partial_l + \partial_i Z^k\Gamma^l_{jk}\partial_l + Z^k\partial_i\Gamma^l_{jk}\partial_l + Z^k\Gamma^m_{jk}\Gamma^l_{im}\partial_l\big) \\ &- \big(\partial_j\partial_i Z^k\partial_k + \partial_i Z^k\Gamma^l_{jk}\partial_l + \partial_j Z^k\Gamma^l_{ik}\partial_l + Z^k\partial_j\Gamma^l_{ik}\partial_l + Z^k\Gamma^m_{ik}\Gamma^l_{jm}\partial_l\big) \\ &= Z^k\big(\partial_i\Gamma^l_{jk} + \Gamma^m_{jk}\Gamma^l_{im} - \partial_j\Gamma^l_{ik} - \Gamma^m_{ik}\Gamma^l_{jm}\big)\partial_l. \end{aligned}$$

This proves[1] that the curvature tensor only depends pointwise on the value of Z, even though it is constructed out of derivatives. The expression in the bracket is called R^l_{ijk}. Because it is often a pain to work with a vector valued function, it is common to define a curvature quadlinear form

$$\mathrm{Rm}(X,Y,Z,W) = g(R(X,Y)Z,W).$$

[1] Can you imagine calculating this without the summation convention?

Clearly the metric depends on W only pointwise, so it makes sense to express this in a chart as

$$\operatorname{Rm}(X,Y,Z,W) = X^iY^jZ^kW^lg_{lm}R^m_{ijk} =: X^iY^jZ^kW^l \operatorname{Rm}_{ijkl}\,.$$

Note, different authors use different conventions about the order of the indices. We follow Lee, whereas Jost and Wikipedia use the order $lkij$. Petersen refuses to choose a side by only ever using $\operatorname{Rm}(\partial_i,\partial_j,\partial_k,\partial_l)$. It is also common to use R for both objects and let the position of the indices distinguish them. On paedological grounds we avoid this.

Example 5.2. The plane (or any euclidean space) in its standard chart has Christoffel coefficients identically equal to zero. Therefore its curvature vanishes at all points.

Example 5.3. In Example 3.29 we calculated the Christoffel symbol for $\mathbb{S}^2$ in stereographic coordinates.

$$\begin{aligned}\frac{2x^1}{\|x\|^2+1} &= -\Gamma^1_{11} = -\Gamma^2_{21} = -\Gamma^2_{12} = \Gamma^1_{22},\\ \frac{2x^2}{\|x\|^2+1} &= \Gamma^2_{11} = -\Gamma^1_{21} = -\Gamma^1_{12} = -\Gamma^2_{22}.\end{aligned}$$

We can use that with the above formula to calculate R^l_{ijk}. We prepare ourselves by calculating the partial derivatives

$$\begin{aligned}\frac{\partial}{\partial x^1}\Gamma^1_{22} &= \frac{-2(x^1)^2+2(x^2)^2+2}{(\|x\|^2+1)^2} & \frac{\partial}{\partial x^2}\Gamma^1_{22} &= \frac{-4x^1x^2}{(\|x\|^2+1)^2}\\ \frac{\partial}{\partial x^1}\Gamma^2_{11} &= \frac{-4x^1x^2}{(\|x\|^2+1)^2} & \frac{\partial}{\partial x^2}\Gamma^2_{11} &= \frac{2(x^1)^2-2(x^2)^2+2}{(\|x\|^2+1)^2}\end{aligned}$$

Many of the curvature coefficients are zero just by definition.

$$R^l_{iik} = \partial_i\Gamma^l_{ik} - \partial_i\Gamma^l_{ik} + \Gamma^m_{ik}\Gamma^l_{im} - \Gamma^m_{ik}\Gamma^l_{im} = 0,$$

so of the sixteen coefficients there are at most eight non-zero entries.

$$\begin{aligned}R^l_{ijk} &= \partial_i\Gamma^l_{jk} - \partial_j\Gamma^l_{ik} + \Gamma^m_{jk}\Gamma^l_{im} - \Gamma^m_{ik}\Gamma^l_{jm}\\ R^1_{121} = -R^1_{211} &= \partial_1\Gamma^1_{21} - \partial_2\Gamma^1_{11} + \Gamma^1_{21}\Gamma^1_{11} + \Gamma^2_{21}\Gamma^1_{12} - \Gamma^1_{11}\Gamma^1_{21} - \Gamma^2_{11}\Gamma^1_{22}\\ &= -\partial_1\Gamma^2_{11} + \partial_2\Gamma^1_{22} + \Gamma^2_{11}\Gamma^1_{22} + \Gamma^1_{22}\Gamma^2_{11} - \Gamma^1_{22}\Gamma^2_{11} - \Gamma^2_{11}\Gamma^1_{22}\\ &= 0.\end{aligned}$$

$$\begin{aligned}R^1_{122} = -R^1_{212} &= \partial_1\Gamma^1_{22} - \partial_2\Gamma^1_{12} + \Gamma^1_{22}\Gamma^1_{11} + \Gamma^2_{22}\Gamma^1_{12} - \Gamma^1_{12}\Gamma^1_{21} - \Gamma^2_{12}\Gamma^1_{22}\\ &= \partial_1\Gamma^1_{22} + \partial_2\Gamma^2_{11} - \Gamma^1_{22}\Gamma^1_{22} + \Gamma^2_{11}\Gamma^2_{11} - \Gamma^2_{11}\Gamma^2_{11} + \Gamma^1_{22}\Gamma^1_{22}\\ &= \partial_1\Gamma^1_{22} + \partial_2\Gamma^2_{11}\\ &= \frac{4}{(\|x\|^2+1)^2}.\end{aligned}$$

$$\begin{aligned} R^2_{121} = -R^2_{211} &= \partial_1\Gamma^2_{21} - \partial_2\Gamma^2_{11} + \Gamma^1_{21}\Gamma^2_{11} + \Gamma^2_{21}\Gamma^2_{12} - \Gamma^1_{11}\Gamma^2_{21} - \Gamma^2_{11}\Gamma^2_{22} \\ &= -\partial_1\Gamma^1_{22} - \partial_2\Gamma^2_{11} - \Gamma^2_{11}\Gamma^2_{11} + \Gamma^1_{22}\Gamma^1_{22} - \Gamma^1_{22}\Gamma^1_{22} + \Gamma^2_{11}\Gamma^2_{11} \\ &= -\partial_1\Gamma^1_{22} - \partial_2\Gamma^2_{11} \\ &= -\frac{4}{(\|x\|^2+1)^2}. \end{aligned}$$

$$\begin{aligned} R^2_{122} = -R^2_{212} &= \partial_1\Gamma^2_{22} - \partial_2\Gamma^2_{12} + \Gamma^1_{22}\Gamma^2_{11} + \Gamma^2_{22}\Gamma^2_{12} - \Gamma^1_{12}\Gamma^2_{21} - \Gamma^2_{12}\Gamma^2_{22} \\ &= -\partial_1\Gamma^2_{11} + \partial_2\Gamma^1_{22} + \Gamma^1_{22}\Gamma^2_{11} + \Gamma^2_{11}\Gamma^1_{22} - \Gamma^2_{11}\Gamma^1_{22} - \Gamma^1_{22}\Gamma^2_{11} \\ &= 0. \end{aligned}$$

We see that many of the coefficients are zero, and the non-zero ones are equal up to a sign.

From Example 3.6 we also have the coefficients of the metric in this chart. In particular, they form a diagonal matrix.

$$g_{ij} = \frac{4}{(\|x\|^2+1)^2}\delta_{ij}.$$

Then the other form of the curvature is

$$\mathrm{Rm}_{ijkl} = g_{lm}R^m_{ijk} = \frac{4}{(\|x\|^2+1)^2}R^l_{ijk}.$$

The non-zero coefficients are

$$\mathrm{Rm}_{1221} = -\mathrm{Rm}_{2121} = -\mathrm{Rm}_{1212} = \mathrm{Rm}_{2112} = \frac{16}{(\|x\|^2+1)^4}.$$

These local expressions show us that the curvature tensor determines n^4 smooth functions Rm_{ijkl}. However some symmetries are apparent already from the definition, such as $R(X,Y)Z = -R(Y,X)Z$. Here are the others

Theorem 5.4 (Symmetries)**. (i)** *Rm is antisymmetric in the first pair and last pair of entries:*

$$\mathrm{Rm}(X,Y,Z,W) = -\mathrm{Rm}(Y,X,Z,W) = -\mathrm{Rm}(X,Y,W,Z).$$

(ii) Rm *is symmetric under the exchange of the first and last pair:*

$$\mathrm{Rm}(X,Y,Z,W) = \mathrm{Rm}(Z,W,X,Y).$$

(iii) *R has the following cyclic symmetry, called the* first *or* algebraic Bianchi identity*:*

$$R(X,Y)Z + R(Z,X)Y + R(Y,Z)X = 0.$$

Proof. (i) seems the logical place to start. We have already noted that antisymmetry in the first pair comes from the definition of R.

$$R(X,Y)Z = \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z.$$

For antisymmetry in the last pair we first compute for $Z = W$ and use metric-compatibility

$$\begin{aligned} g(R(X,Y)Z,Z) &= g(\nabla_X\nabla_Y Z, Z) - g(\nabla_Y\nabla_X Z, Z) - g(\nabla_{[X,Y]}Z, Z) \\ &= X(g(\nabla_Y Z, Z)) - g(\nabla_Y Z, \nabla_X Z) - Y(g(\nabla_X Z, Z)) + g(\nabla_X Z, \nabla_Y Z) \\ &\quad - \frac{1}{2}[X,Y](g(Z,Z)) \\ &= X(g(\nabla_Y Z, Z)) - Y(g(\nabla_X Z, Z)) - \frac{1}{2}[X,Y](g(Z,Z)) \\ &= X(\frac{1}{2}Y(g(Z,Z))) - Y(\frac{1}{2}X(g(Z,Z))) - \frac{1}{2}[X,Y](g(Z,Z)) = 0. \end{aligned}$$

Any bilinear function that is zero on (Z,Z) is antisymmetric:

$$\begin{aligned} 0 &= g(R(X,Y)Z + W, Z + W) - g(R(X,Y)Z - W, Z - W) \\ &= 2g(R(X,Y)Z, W) + 2g(R(X,Y)W, Z). \end{aligned}$$

Next we prove (iii). The follows from a long calculation, but one that can be shortened using the following piece of notation from Petersen:

$$\mathcal{S}T(X,Y,Z) = T(X,Y,Z) + T(Z,X,Y) + T(Y,Z,X).$$

$$\begin{aligned} \mathcal{S}R(X,Y)Z &= \mathcal{S}\nabla_X\nabla_Y Z - \mathcal{S}\nabla_Y\nabla_X Z - \mathcal{S}\nabla_{[X,Y]}Z \\ &= \mathcal{S}\nabla_Z\nabla_X Y - \mathcal{S}\nabla_Z\nabla_Y X - \mathcal{S}\nabla_{[X,Y]}Z \\ &= \mathcal{S}\nabla_Z(\nabla_X Y - \nabla_Y X) - \mathcal{S}\nabla_{[X,Y]}Z \\ &= \mathcal{S}\big(\nabla_Z[X,Y] - \nabla_{[X,Y]}Z\big) \\ &= \mathcal{S}[Z,[X,Y]], \end{aligned}$$

using twice that ∇ is torsion-free. This expression is always zero, a fact known as the *Jacobi identity*, which is easily proved using a chart:

$$\mathcal{S}[Z,[X,Y]] = [Z,[X,Y]] + [X,[Y,Z]] + [Y,[Z,X]] = 0.$$

Now (ii) follows from (i) and (iii)

$$\begin{aligned} \mathrm{Rm}(X,Y,Z,W) &= -\mathrm{Rm}(Z,X,Y,W) - \mathrm{Rm}(Y,Z,X,W) \\ &= \mathrm{Rm}(Z,X,W,Y) + \mathrm{Rm}(Y,Z,W,X) \\ &= \mathrm{Rm}(W,Z,X,Y) - \mathrm{Rm}(X,W,Z,Y) - \mathrm{Rm}(W,Y,Z,X) - \mathrm{Rm}(Z,W,Y,X) \\ &= 2\,\mathrm{Rm}(Z,W,X,Y) + \mathrm{Rm}(X,W,Y,Z) + \mathrm{Rm}(W,Y,X,Z) \\ &= 2\,\mathrm{Rm}(Z,W,X,Y) - \mathrm{Rm}(X,Y,Z,W). \end{aligned}$$

□

Exercise 5.5. Show that these symmetries restrict the number of independent coefficients of Rm to $n^2(n^2-1)/12$. In particular, for $n = 2$ there is essentially only one coefficient.

The algebraic Bianchi identity was first written down by Ricci. However it is so named because it looks similar to a cyclic identity discovered by Bianchi. It's general form requires additional definitions of a kind we have avoided, so we give a special form that is suitable to applications.

Theorem 5.6. *For any $p \in M$ take the normal chart centered at p such that $g_{ij}(p) = \delta_{ij}$. The* second *or* differential Bianchi identity *states that at p:*

$$\partial_m \operatorname{Rm}_{ijkl} + \partial_k \operatorname{Rm}_{ijlm} + \partial_l \operatorname{Rm}_{ijmk} = 0.$$

Proof. In the normal chart at p this point is the origin. Due to Lemma 4.34 we know in the normal chart at p that $g_{ij}(p) = \delta_{ij}$ and $\partial_k g_{ij}(p) = 0$. In the proof of that lemma it was shown, and it follows easily from $\partial_k g_{ij}(0) = 0$, that $\Gamma^k_{ij}(0) = 0$. Hence

$$\begin{aligned}(\partial_m \operatorname{Rm}_{ijkl})(0) &= (\partial_m(g_{ln}R^n_{ijk}))(0) = 0 + \delta_{ln}(\partial_m R^n_{ijk})(0) = \partial_m R^l_{ijk}(0)\\ &= \partial_m\big(\partial_i\Gamma^l_{jk} - \partial_j\Gamma^l_{ik} + \Gamma^m_{jk}\Gamma^l_{im} - \Gamma^m_{ik}\Gamma^l_{jm}\big)(0)\\ &= \partial_m\partial_i\Gamma^l_{jk}(0) - \partial_m\partial_j\Gamma^l_{ik}(0) + (0+0) - (0+0). \qquad (5.7)\end{aligned}$$

Taking the cyclic permutations of ijm proves

$$(\partial_m \operatorname{Rm}_{ijkl})(0) + (\partial_i \operatorname{Rm}_{jmkl})(0) + (\partial_j \operatorname{Rm}_{mikl})(0) = 0.$$

And swapping the first and last pair gives the identity as stated in the theorem. □

We motivated the introduction of the curvature tensor by asking whether a space was locally isometric to euclidean space, specifically whether there existed a parallel vector field in a neighbourhood of a point. Historically this question was approached through the lens of coordinate transformation: does there exist a coordinate transformation that makes the metric coefficients constant and equal to δ_{ij}? The relevance of normal coordinates, where $g_{ij}(p) = \delta_{ij} + 0 + \mathcal{O}(\|p\|^2)$, to the question now seems obvious. What Riemann found was that there was an obstacle in the second order of the Taylor expansion that could not be removed. To see the relation between our definition and Riemann's observation, start with the formula for R^l_{ijk} in terms of the Christoffel coefficients and substitute in the expression for them in terms of the metric coefficients. Suppose you have a chart where $g_{ij}(p) = \delta_{ij} + 0 + \mathcal{O}(\|p\|^2)$. The normal chart has this property, but there may be others.

$$\begin{aligned}\partial_i\Gamma^l_{jk} &= \partial_i g^{n\,l}\big(\partial_k g_{j\,n} + \partial_j g_{k\,n} - \partial_n g_{k\,j}\big) + g^{n\,l}\partial_i\big(\partial_k g_{j\,n} + \partial_j g_{k\,n} - \partial_n g_{k\,j}\big)\\ \partial_i\Gamma^l_{jk}(0) &= 0 + \partial_i\partial_k g_{j\,l} + \partial_i\partial_j g_{k\,l} - \partial_i\partial_l g_{k\,j}\\ \partial_j\Gamma^l_{ik}(0) &= \partial_j\partial_k g_{i\,l} + \partial_j\partial_i g_{k\,l} - \partial_j\partial_l g_{k\,i}\\ R^l_{ijk} &= \partial_i\Gamma^l_{jk} - \partial_j\Gamma^l_{ik} + \Gamma^m_{jk}\Gamma^l_{im} - \Gamma^m_{ik}\Gamma^l_{jm}\\ R^l_{ijk}(0) &= \partial_i\Gamma^l_{jk}(0) - \partial_j\Gamma^l_{ik}(0) + 0 - 0\\ &= \partial_i\partial_k g_{j\,l}(0) - \partial_i\partial_l g_{k\,j}(0) - \partial_j\partial_k g_{i\,l}(0) + \partial_j\partial_l g_{k\,i}(0).\end{aligned}$$

If in this coordinate chart all the second derivatives of g also vanished at some point, then the right hand side would be zero. This implies $R(X,Y)Z = 0$ for all vectors $X, Y, Z \in T_pM$. But curvature is defined independent of coordinates charts. If the curvature tensor is non-zero at some point p in some directions, logically it is impossible in any chart for the Taylor series of g_{ij} at p to additionally vanish in the second order. In particular curvature is an obstruction to being locally euclidean.

As we saw in Theorem 5.4, the numerous symmetries of R mean that there is a lot of redundancy in its coefficients. It makes sense therefore to ask if there is a way to distil the information of the curvature tensor into a simpler object. We provide two such simplifications now, and will look at a third in the next section.

Definition 5.8. *For every point $p \in M$ and vectors $Y, Z \in T_pM$, we consider the linear map $X \mapsto R(X,Y)Z$ from T_pM to itself. The* Ricci *curvature* $\mathrm{Ric}(Y,Z)$ *is the trace of this map. It is bilinear in Y, Z so can be expressed nicely using coefficients*

$$\mathrm{Ric}(Y,Z) := Y^j Z^k \,\mathrm{Ric}_{jk} = Y^j Z^k \delta^i_l R^l_{ijk} = Y^j Z^k g^{im}\,\mathrm{Rm}_{ijkm}\,.$$

Likewise the scalar curvature *S is the trace (in the sense of bilinear forms) of the Ricci curvature with respect to the metric:*

$$S = \mathrm{tr}_g \,\mathrm{Ric} = g^{jk}\,\mathrm{Ric}_{jk} = g^{jk} g^{im}\,\mathrm{Rm}_{ijkm}\,.$$

It may seem more natural to take the trace of $Z \mapsto R(X,Y)Z$. However this is zero because Rm is antisymmetric in the last pair. Likewise antisymmetry of R in X, Y means taking the trace of $X \mapsto R(Y,X)Z$ just gives a negative sign. The Ricci curvature is in fact a symmetric bilinear form:

$$\mathrm{Ric}_{kj} = g^{im}\,\mathrm{Rm}_{ikjm} = g^{im}\,\mathrm{Rm}_{jmik} = g^{mi}\,\mathrm{Rm}_{mjki} = g^{im}\,\mathrm{Rm}_{ijkm} = \mathrm{Ric}_{jk}\,.$$

Example 5.9. We can continue our example of the sphere $\mathbb{S}^2$ in stereographic coordinates. In Example 5.3 we computed the coefficients R^l_{ijk} in the chart U_N. Most were zero. Therefore the coefficients of the Ricci tensor in this chart are

$$\begin{aligned}
\mathrm{Ric}_{jk} &= \delta^i_l R^l_{ijk} = R^1_{1jk} + R^2_{2jk} \\
\mathrm{Ric}_{11} &= R^1_{111} + R^2_{211} = R^2_{211} = \frac{4}{(\|x\|^2+1)^2} \\
\mathrm{Ric}_{12} &= R^1_{112} + R^2_{212} = 0 \\
\mathrm{Ric}_{21} &= R^1_{121} + R^2_{221} = 0 \\
\mathrm{Ric}_{22} &= R^1_{122} + R^2_{222} = R^1_{122} = \frac{4}{(\|x\|^2+1)^2}.
\end{aligned}$$

As expected, this is a symmetric matrix.

For the scalar curvature we need the the inverse of the matrix of the metric

$$g^{ij} = \frac{(\|x\|^2+1)^2}{4}\delta^{ij}.$$

Hence

$$S = g^{jk}\,\mathrm{Ric}_{jk} = \frac{(\|x\|^2+1)^2}{4}\,(\mathrm{Ric}_{11} + \mathrm{Ric}_{22}) = 2.$$

Not unreasonably, the scalar curvature for the sphere is at every point 2.

We will not go deeply into the theory of Ricci and scalar curvature, but we will mention some special cases of interest. Spaces with $\mathrm{Ric} \equiv 0$ are called *Ricci-flat*. A slightly more general class of Riemannian manifolds are *Einstein* manifolds. These have the property that $\mathrm{Ric} = \lambda g$ for a function $\lambda : M \to \mathbb{R}$. Since the Ricci curvature must be a symmetric bilinear form, this is more-or-less the simplest form it could take. Taking trace of both sides shows that for Einstein manifolds,

$$S = g^{jk}\,\mathrm{Ric}_{jk} = \lambda g^{jk} g_{jk} = \lambda \sum_k \delta^k_k = \lambda \dim M.$$

Example 5.10. Observe that the sphere is an Einstein manifold with $\lambda \equiv 1$ as

$$\mathrm{Ric}_{jk} = \frac{4}{(\|x\|^2+1)^2}\delta_{ij} = g_{ij}.$$

We also see that its scalar curvature is $S = \lambda \dim \mathbb{S}^2 = 1 \times 2 = 2$.

These are named for Einstein because the equation for the curvature of space-time in the theory of general relativity is

$$\mathrm{Ric}_{ij} - \frac{1}{2} S g_{ij} = T_{ij},$$

where the right hand side is a function representing matter-energy. If you allow on the left hand side an additional 'cosmological constant'

$$\mathrm{Ric}_{ij} - \frac{1}{2} S g_{ij} + \Lambda g_{ij} = T_{ij},$$

then Einstein manifolds are models of a vacuum universe (no matter-energy). Einstein originally published his theory without a cosmological constant. At the time it was thought that the universe was static and eternal, and in a subsequent publication he argued for Λ to permit this. A decade later, the observations of distant galaxies by Hubble showed that the universe was expanding. He would call this his "biggest blunder", as trusting the simplicity of his original derivation would have meant another successful prediction of the theory. It seems a little mean naming this class of manifolds after a man's biggest blunder.

Theorem 5.11 (Schur)**.** *On a connected Einstein manifold with dimension three or greater, the scalar curvature is constant.*

Proof. The key to this proof is the differential Bianchi identity 5.6. In normal coordinates at a point p

$$(\partial_m \mathrm{Ric}_{jk})(p) = (\partial_m (g^{il} \mathrm{Rm}_{ijkl}))(p) = 0 + \delta^{il}(\partial_m(\mathrm{Rm}_{ijkl}))(p).$$

On the other hand

$$(\partial_m(\lambda g_{jk}))(p) = (\partial_m \lambda)(p)\delta_{jk} + 0.$$

Putting this into the Bianchi identity gives (leaving evaluation at p implicit)

$$\begin{aligned}
0 &= \partial_m \mathrm{Rm}_{ijkl} - \partial_k \mathrm{Rm}_{ijml} + \partial_l \mathrm{Rm}_{ijmk} \\
0 &= \delta^{il}\partial_m \mathrm{Rm}_{ijkl} - \delta^{il}\partial_k \mathrm{Rm}_{ijml} + \delta^{il}\partial_l \mathrm{Rm}_{ijmk} \\
&= \partial_m \lambda\, \delta_{jk} - \partial_k \lambda\, \delta_{jm} + \delta^{il}\partial_l \mathrm{Rm}_{ijmk} \\
0 &= \delta^{jk}\partial_m \lambda\, \delta_{jk} - \delta^{jk}\partial_k \lambda\, \delta_{jm} + \delta^{jk}\delta^{il}\partial_l \mathrm{Rm}_{ijmk} \\
&= \partial_m \lambda\, \dim M - \partial_m \lambda - \delta^{il}\delta^{jk}\partial_l \mathrm{Rm}_{jimk} \\
&= \partial_m \lambda\, (\dim M - 1) - \delta^{il}\partial_l \lambda\, \delta_{im} \\
&= \partial_m \lambda\, (\dim M - 2).
\end{aligned}$$

If $\dim M > 2$ then this forces λ to have zero derivative in every direction. Morever, we can do this for every point. Therefore $\lambda = (\dim M)^{-1} S$ is constant. □

5.2 Hypersurfaces

A hypersurface is the embedding of an n-dimensional manifold M in an $(n+1)$-dimensional manifold N (codimension one). Additionally, assume that both a Riemannian manifolds and that the embedding is Riemannian. Alternatively, you may begin with a Riemannian manifold N and any manifold M and then put the pullback metric on M, which will make the embedding Riemannian. Recall Definition 3.54 and Theorem 3.55. They explain how the Levi-Civita connection $\nabla^\top$ of M can be calculated using the Levi-Civita connection ∇^N; essentially $\nabla^\top$ is the projection of ∇^N.

We can also ask what information is lost by this projection.

Definition 5.12. *The* second fundamental form *of M in N is the function*

$$\mathrm{I\!I}(X,Y) = \nabla^N_X Y - \nabla^\top_X Y.$$

This is a function from tangent vector fields on M to a normal vector field on M.

In the case that M is a hypersurface in N, there is an up-to-sign unique unit normal vector field ν of M.[2] *In this case we also consider the second fundamental form to be*

$$h(X,Y) = g(\mathrm{I\!I}(X,Y),\nu) = g(\nabla^N_X Y, \nu).$$

That this definition aligns with the definition in Section 1.5 is essentially the Meusnier's theorem 1.25. For $\mathbb{R}^3$ the metric is the dot product and the covariant derivative is just the usual directional derivative. The theorem tells us for a arc-length parameterised curve that $h(\alpha',\alpha')$ is equal to the normal curvature $\alpha''\cdot\nu$. In $\mathbb{R}^3$ it was clear by definition that h was a symmetric C^∞-bilinear form. This property generalises.

Theorem 5.13 (Codazzi-Mainardi)**.** *The second fundamental form $\mathrm{I\!I}$ is a symmetric C^∞-bilinear function.*

Proof. Consider the antisymmetric part of $\mathrm{I\!I}$. For tangent vector fields X, Y to M, we compute

$$\mathrm{I\!I}(X,Y) - \mathrm{I\!I}(Y,X) = \nabla^N_X Y - \nabla^\top_X Y - \nabla^N_Y X + \nabla^\top_Y X = T^N(X,Y) - T^\top(X,Y) = 0,$$

where T^N and $T^\top$ are torsion. But Levi-Civita connections are torsion-free. This proves the symmetry.

By definition covariant derivatives are C^∞-linear in the direction. Hence $\mathrm{I\!I}(X,Y)$ is C^∞-linear in X. But by symmetry it is also C^∞-linear in Y. □

Notice that the Gauss formula only applies to tangent vector fields of M. We can also ask about the derivative of the normal vector field ν. Observe that because ν is unit-length, for any tangent vector $X \in T_pM$

$$0 = X(g(\nu,\nu)) = 2g(\nu, \nabla^N_X \nu)$$

[2] If M is orientable this field is global. If M is non-orientable then this can only be chosen locally.

shows that $\nabla^N_X \nu$ is tangent to M. In fact

$$0 = X(g(Y,\nu)) = g(\nabla^N_X Y, \nu) + g(Y, \nabla^N_X \nu) = g(\nabla^\top_X Y + \mathrm{I\!I}(X,Y), \nu) + g(Y, \nabla^N_X \nu)$$
$$\Rightarrow \quad g(Y, \nabla^N_X \nu) = -h(X,Y)$$

The above is called the *Weingarten formula* and is the analogue of the working following Exercise 1.26. It tells us that for a hypersurface, the covariant derivative of any vector field in N in a tangent direction of M can be calculated with $\nabla^\top$ and h alone.

From these formulae follows a particularly nice formula relating the Riemann curvature tensors of M and N. Originally the Levi-Civita connection was simply defined to the the tangent connection, so this formula was called the Gauss formula.

$$\begin{aligned}
\mathrm{Rm}^N(X,Y,Z,W) &= g(R^N(X,Y)Z,W) = g(\nabla^N_X \nabla^N_Y Z - \nabla^N_Y \nabla^N_X Z - \nabla^N_{[X,Y]} Z, W) \\
&= g(\nabla^N_X(\nabla^\top_Y Z + h(Y,Z)\nu) - \nabla^N_Y(\nabla^\top_X Z + h(X,Z)\nu) - \nabla^\top_{[X,Y]} Z, W) \\
&= g(\nabla^\top_X \nabla^\top_Y Z + \nabla^N_X(h(Y,Z)\nu) - \nabla^\top_Y \nabla^\top_X Z - \nabla^N_Y(h(X,Z)\nu) - \nabla^\top_{[X,Y]} Z, W) \\
&= g(R^M(X,Y)Z + \nabla^N_X(h(Y,Z)\nu) - \nabla^N_Y(h(X,Z)\nu), W) \\
&= \mathrm{Rm}^M(X,Y,Z,W) + g(X(h(Y,Z))\nu + h(Y,Z)\nabla^N_X \nu, W) \\
&\quad - g(Y(h(X,Z))\nu + h(X,Z)\nabla^N_Y \nu, W) \\
&= \mathrm{Rm}^M(X,Y,Z,W) + h(Y,Z)g(\nabla^N_X \nu, W) - h(X,Z)g(\nabla^N_Y \nu, W) \\
&= \mathrm{Rm}^M(X,Y,Z,W) - h(Y,Z)h(X,W) + h(X,Z)h(Y,W).
\end{aligned}$$

Because for euclidean space the Riemann curvature tensor vanishes, this gives us a relation between the curvature tensor of M and the second fundamental form.

Theorem 5.14 (Theorema Egregium)**.** *Let M be a Riemannian embedded surface in $\mathbb{R}^3$. Let $X, Y \in T_pM$ be orthonormal vectors. The Gauss curvature K is related to the curvature tensor by*

$$K = \mathrm{Rm}(X,Y,Y,X).$$

Therefore the Gauss curvature of a surface is an isometry invariant.

There exists a modification of this formula for arbitrary tangent vectors, but it amounts to applying Gram-Schmidt orthogonalisation to the vectors X, Y . Therefore we work exclusively with this more elegant form.

Proof. If $X, Y \in T_pM$ are orthonormal, then we can use them as a basis of T_pM. With respect to this basis

$$h = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix} = \begin{pmatrix} h(X,X) & h(X,Y) \\ h(Y,X) & h(Y,Y) \end{pmatrix}.$$

Using Lemma 1.34 and the Gauss formula for curvature gives us therefore that

$$K = h_{11}h_{22} - h_{12}^2 = h(Y,Y)h(X,X) - h(X,Y)h(Y,X) = \mathrm{Rm}^M(Y,X,X,Y). \qquad \square$$

Example 5.15. Again we consider the example of the sphere $\mathbb{S}^2$. We know the Riemann curvature already, so we should be able to use that to calculate the Gauss curvature. First we need orthonormal vectors at every point of the chart U_N. The coordinate vectors are orthogonal to one another, but not unit length. Therefore take

$$X = \frac{\|x\|^2+1}{2}\partial_1, \qquad Y = \frac{\|x\|^2+1}{2}\partial_2.$$

Then

$$K = \mathrm{Rm}(X,Y,Y,X) = \left(\frac{\|x\|^2+1}{2}\right)^4 \mathrm{Rm}_{1221} = 1.$$

As we observed in Example 1.31, the normal curvature of a sphere at every point and in every direction is the inverse of its radius, here 1. Hence the principal curvatures are 1 and directly from Definition 1.33 we see that the Gauss curvature is 1, in agreement with the above calculation.

Definition 5.16. *If M is a 2-dimensional Riemannian manifold, we define the Gauss curvature $K(p) = \mathrm{Rm}(X,Y,Y,X)$ for any orthonormal basis X, Y of T_pM.*

Exercise 5.17. Implicit in this definition is the claim that this quantity is independent of the choice of orthonormal basis of T_pM. Proof this claim.

5.3 Sectional Curvature

The Ricci and scalar curvatures are natural simplifications of the curvature tensor and have nice properties, but we are yet to see any geometric intuition for these so-called curvatures. In this section we finally address this question.

In Chapter 1 we defined the curvature of a surfaces by reference to the curvature of curves in that surface. In particular the normal curvature was important, which we now understand is the curvature of a geodesic. Likewise we can define a type of curvature using surfaces within our manifold. These surfaces should be special in some way so that their curvature reflects the curvature of the manifold.

Definition 5.18. *For any point $p \in M$ let us use the normal chart at p. Given a pair of orthonormal vectors $X, Y \in T_pM$, they span a plane in the chart, called the* plane section *P. The* sectional curvature *$K(P) = K(X, Y)$ of M at p is defined to be the Gauss curvature at p of the plane section.*

Example 5.19. Consider a two dimensional manifold. Then in the chart there is only one possible plane, namely the chart itself. In this case the definitions are somewhat trivial and the sectional curvature is just the Gauss curvature of the manifold from Definition 5.16.

In general it is important to use normal coordinates to define the plane section. In some sense, this plane is constructed out of geodesics. We first show that the second fundamental form of the plane section in M is zero at p. Choose any vector $v \in T_pP$. Let γ_v be the geodesic through p in the direction of v. The definition of the normal chart is that this is a ray, hence it lies in P. By the Gauss formula,

$$0 = \nabla^M_{\gamma'}\gamma' = \nabla^P_{\gamma'}\gamma' + \mathrm{I\!I}(\gamma', \gamma').$$

The tangent and normal directions are linearly independent, so both terms on the right must vanish. In particular $0 = \mathrm{I\!I}(\gamma'(0), \gamma'(0)) = \mathrm{I\!I}(v, v)$ for all $v \in T_pM$. Since $\mathrm{I\!I}$ is symmetric it must vanish at p. Together with Theorema Egregium 5.14 and the Gauss formula for curvature, we have proved

Theorem 5.20. *If $X, Y \in T_pM$ are orthonormal then the sectional curvature of M at p is*

$$K(X, Y) = \mathrm{Rm}^P(X, Y, Y, X) = \mathrm{Rm}^M(X, Y, Y, X).$$

It may seem as though the sectional curvature is just another simplification of the full curvature tensor. But this is not the case.

Theorem 5.21. *The sectional curvatures uniquely determine the curvature tensor. Moreover if at some point $p \in M$ the sectional curvature is constant $K(X, Y) = \kappa$ for all orthonormal $X, Y \in T_pM$ then the curvature tensor is given by*

$$\mathrm{Rm}(X, Y, Z, W) = \kappa\big(g(X, W)g(Y, Z) - g(X, Z)g(Y, W)\big).$$

Proof. Suppose two curvature tensors have the same sectional curvatures at a point p. Then their difference $\mathcal{R}$ is a quadlinear map on T_pM with the same symmetries as a curvature tensor.

It is sufficient to prove $\mathcal{R}(X,Y,Z,W)=0$ for X,Y,Z,W elements of an orthonormal basis for T_pM. We now use the symmetries:

$$\begin{aligned}0&=\mathcal{R}(X+W,Y,Y,X+W)\\&=\mathcal{R}(X,Y,Y,X)+\mathcal{R}(X,Y,Y,W)+\mathcal{R}(W,Y,Y,X)+\mathcal{R}(W,Y,Y,X)\\&=0+2\mathcal{R}(X,Y,Y,W)+0,\end{aligned}$$

and

$$\begin{aligned}0&=\mathcal{R}(X,Y+Z,Y+Z,W)\\&=\mathcal{R}(X,Y,Y,W)+\mathcal{R}(X,Y,Z,W)+\mathcal{R}(X,Z,Y,W)+\mathcal{R}(X,Z,Z,W)\\&=0+\mathcal{R}(X,Y,Z,W)+\mathcal{R}(X,Z,Y,W)+0.\end{aligned}$$

In other words, in addition to being antisymmetric in the first and last pair, $\mathcal{R}$ is also antisymmetric in the middle pair. Finally we apply the algebraic Bianchi identity:

$$\begin{aligned}0&=\mathcal{R}(X,Y,Z,W)+\mathcal{R}(Z,X,Y,W)+\mathcal{R}(Y,Z,X,W)\\&=\mathcal{R}(X,Y,Z,W)-\mathcal{R}(X,Z,Y,W)-\mathcal{R}(Y,X,Z,W)\\&=\mathcal{R}(X,Y,Z,W)+\mathcal{R}(X,Y,Z,W)+\mathcal{R}(X,Y,Z,W).\end{aligned}$$

Hence $\mathcal{R}(X,Y,Z,W)=0$ as required.

For the second claim, we observe that if a curvature tensor of the given form has constant sectional curvature then from the preceding argument it must be the unique such curvature tensor. For orthonormal $X,Y\in T_pM$ we have

$$K(X,Y)=\operatorname{Rm}(X,Y,Y,X)=\kappa\big(g(X,X)g(Y,Y)-g(X,Y)g(Y,X)\big)=\kappa.$$

□

Example 5.22. We see now that the curvature of the sphere $\mathbb{S}^2$ has exactly this form. The Gauss curvature at every point is 1. Then

$$\operatorname{Rm}_{ijkl}=\operatorname{Rm}(\partial_i,\partial_j,\partial_k,\partial_l)=1\big(g_{il}g_{jk}-g_{ik}g_{jl}\big)=\left(\frac{4}{(\|x\|^2+1)^2}\right)^2\big(\delta_{il}\delta_{jk}-\delta_{ik}\delta_{jl}\big).$$

We had in Theorem 5.11 that Einstein manifolds, which have a special form of the Ricci curvature, have constant scalar curvature. There is a similar result in the case of sectional curvature.

Theorem 5.23 (Schur)**.** *Suppose that a connected manifold of dimension three or more has pointwise constant sectional curvature. This means there is a function $\kappa: M\to\mathbb{R}$ such that for all orthonormal $X,Y\in T_pM$ we have $K(X,Y)=\kappa(p)$. Then κ is constant.*

Proof. At any point $p\in M$ we can use a normal chart. In this chart

$$\partial_m \operatorname{Rm}(\partial_i,\partial_j,\partial_k,\partial_l)|_p=\partial_m(\kappa(g_{il}g_{jk}-g_{ik}g_{jl}))|_p=\partial_m\kappa(\delta_{il}\delta_{jk}-\delta_{ik}\delta_{jl})$$

The differential Bianchi identity says

$$\begin{aligned} 0 &= \partial_m \operatorname{Rm}_{ijkl} + \partial_k \operatorname{Rm}_{ijlm} + \partial_l \operatorname{Rm}_{ijmk} \\ &= \partial_m \kappa(\delta_{il}\delta_{jk} - \delta_{ik}\delta_{jl}) + \partial_k \kappa(\delta_{im}\delta_{jl} - \delta_{il}\delta_{jm}) + \partial_l \kappa(\delta_{ik}\delta_{jm} - \delta_{im}\delta_{jk}). \end{aligned}$$

In particular, choose $l = i$, $k = j$, and m, i, j distinct. It is possible for three indices to be distinct because the dimension is at least three.

$$0 = \partial_m \kappa(1 - 0) + \partial_k \kappa(0 - 0) + \partial_l \kappa(0 - 0) = \partial_m \kappa.$$

Thus κ has zero derivative at p, and hence at every point. It must be constant. □

Spaces with constant sectional curvature are called *space forms*. In every dimension the space forms are the euclidean space, the sphere (with scalings), and hyperbolic plane (with scalings), this classification is due to Killing and Hopf. Thus these three spaces, which have been the main focus of our examples, are in terms of Riemannian geometry the nicest spaces. We have in a previous section mentioned the study of manifolds with special curvature. Another direction of research is to impose a bound on the curvature. For example, a theorem of Myers states that if a Riemannian manifold is complete as a metric space and the infimum of its sectional curvatures is positive, then it is compact. Or a result of Synge says that a compact orientable even-dimensional Riemannian manifold with positive sectional curvatures must be simple connected.

Finally, both theorems of Schur rely on the dimension being three or greater. This is not a limitation of the proof: Riemann surfaces really are special. There is only one intrinsic curvature for them, the Gauss curvature, and integral of the Gauss curvature over the whole manifold is closely connected with its topology (the Gauss-Bonnet theorem). Moreover every Riemann surface is conformally equivalent to a Riemann surface with constant curvature, which then are the three space forms. More can be learnt about the theory of Riemann surfaces in Complex Analysis II (Funktiontheorie II).

Appendix A

Literature

All textbooks that I know of approach differential geometry in the following sequence: the classical subject of curves and surfaces, the general theory of manifolds, Riemannian manifolds. In fact most only treat one of the three topics. It was somewhat of a challenge therefore to produce a script that opened with foundational material, covered the least amount of manifold theory possible, and still was able get to interesting results in Riemannian geometry. I'm writing this before semester begins, so there is no guarantee we actually make it all the way through!

Do Carmo's *Differential Geometry of Curves and Surfaces* (1976, 2016) is a classic. It has great illustrations and spends a lot of time examining special classes of surfaces. This gives the reader a good grounding in the sorts of problems that might be interesting in higher dimensions. Our Chapter 1 does not follow do Carmo as closely as it otherwise might because I wanted to motivate the curvature of curves through the osculating circle. The approach to normal curvature and Meusnier's theorem is very much in line with do Carmo however.

Chapter 2 is a strange beast. Some amount of manifold theory is unavoidable to talk about Riemannian manifolds. But I didn't want to just turn this course into "Analysis III Lite" and I wanted students who have already taken Analysis III to get something new out of it. Therefore I tried to write down a fresh approach that centers charts as the primary object of manifold theory. This is somewhat inspired by the physics approach to general relativity, where every observer has their own chart. I have no good references to give.

There after we move into what is a more-or-less standard approach to Riemannian geometry. I relied most heavily on John M Lee's *Introduction to Riemannian Manifolds* (second edition, 2018). To name but one example, we followed his convention on the indices of the curvature coefficients. More generally, we followed his example by treating geodesics before curvature. I recommend this textbook to a student looking to reinforce their knowledge because of its strong explanations and willingness to use geometric intuitions.

The second textbook I leaned on for the later chapters was Petersen's *Riemannian Geometry* (2005). I was particularly impressed with his handling of the symmetries of the Riemannian curvature. Prof Schmidt's lecture notes from when he last ran this course, which I may have shared with you since they are in German, follows the path set by Petersen. However I would recommend this book only for a student who is already confident with Riemannian geometry, because it is rather brisk and it shies away from discussions. This means however it has time

for further explorations (curvature bounds and symmetric spaces). A book that is similar in this respect but with a different set of extension topics (harmonic maps and Floer homology) is Jost's *Riemannian Geometry and Geometric Analysis* (2017).

Speaking of geometric analysis, I mentioned in Chapter 4 geometric flows as an alternative method of straightening curves. I would recommend interested students check out Andrews et al *Extrinsic Geometric Flows* (2020). Already with the knowledge of curves and their curvatures from this course the second chapter of that book, on the curve shortening flow, is very readable.

Likewise I have a special spot in my heart for Sharpe's *Differential geometry: Cartan's generalization of Klein's Erlangen program*. It really represents a completely different approach to geometry. Klein's Erlangen program was an attempt to unify all the then-known geometries (euclidean, hyperbolic, elliptic, projective, etc) into one framework. The idea worked, but it never became as popular as the tensor calculus of the Italians, probably due to the influence of general relativity. What is interesting from the perspective of modern differential geometry is that many of the 'old ideas' that have been dropped or relegated to corollaries are first-class ideas in this other approach. For example, moving a tangent plane along a surface using euclidean motions is a very natural operations from this point of view, which is why we mentioned this book in Chapter 3.

Appendix B

Linear Algebra

Cross product

Consider $\mathbb{R}^n$ with the dot product. With this we can determine the lengths of vectors and the angle between them. In fact the relation $a \cdot b = \|a\| \, \|b\| \cos\theta$ says that the two are equivalent information. The reason to use the dot product over simply the length and angle information is because it is very useful to encode this information into a bilinear operation. Likewise, in $\mathbb{R}^3$ given two non-parallel vectors in order, there is a third perpendicular vector that completes the oriented basis. It is useful to have an operation that goes from a pair of vectors to the perpendicular complement, but even more useful if this operation has nice algebraic properties.

Here is a construction that leads to the standard cross product. It must be anticommutative to encode the orientation. We want it to be bilinear, so it is sufficient to construct it on unit vectors. Note that negating a vector changes the orientation so the output should also be negated to preserve the orientation, and this is compatible with bilinearity. Finally, it should be rotationally preserved. With these stimulations, we must choose an odd function $f(\theta)$ that determines

$$\mathbf{i} \times (\cos\theta \, \mathbf{i} + \sin\theta \, \mathbf{j}) = f(\theta)\mathbf{k}.$$

By anticommutativity, $\mathbf{i} \times \mathbf{i} = 0$. If the cross product is to be bilinear, then

$$f(\theta)\mathbf{k} = \sin\theta \, (\mathbf{i} \times \mathbf{j}) = \sin\theta \, f(\pi/2)\mathbf{k}.$$

The obvious choice is to set $f(\pi/2) = 1$ so that $f(\theta) = \sin\theta$.

We see that the cross product is determined essentially by the relation $\mathbf{i} \times \mathbf{j} = \mathbf{k}$ and bilinearity. This can be used to calculate any cross product and is called the 'algebraic method'. Alternatively, we have $\|a \times b\| = \|a\| \, \|b\| \, |\sin\theta|$. We can interpret the right hand side as the area of the parallelogram spanned by a and b. Together with the perpendicularity, this information also determines the cross product. It is called the 'geometric method'. This is the reason that cross products often appear in surface area formulas. For example, a parameterised surface $\Phi : U \subseteq \mathbb{R}^2 \to \mathbb{R}^3$ has a surface area of

$$\iint_U \|\partial_u \Phi \times \partial_v \Phi\| \, du \, dv.$$

Finally, the computation method I use the most is the 'determinant method'. Given $a = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ and $b = b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}$ take the following determinant

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = (a_2b_3 - a_3b_2)\mathbf{i} + (a_3b_1 - a_1b_3)\mathbf{j} + (a_1b_2 - a_2b_1)\mathbf{k}.$$

Although it is an abomination that mixes vectors and scalars in one matrix, it gives the correct result.

Linear Atlases

In this section we offer a different perspective on vector space theory. It is well known that every finitely generated vector space over $\mathbb{K}$ is isomorphic to $\mathbb{K}^n$ for a single $n \in \mathbb{N}^+$ called the dimension. This is often summarised as 'there is only one vector space per (finite) dimension'. One proves this by construction of an ordered basis, which is equivalent to a linear isomorphism to $\mathbb{K}^n$ via

$$\begin{aligned} (v_1, \ldots, v_n) \in V \quad &\mapsto \quad \phi(v_i) := e_i \\ \phi : V \to \mathbb{K}^n \quad &\mapsto \quad (\phi^{-1}(e_1), \ldots, \phi^{-1}(e_n)), \end{aligned}$$

where $(e_1, \ldots, e_n)$ is the standard basis of $\mathbb{K}^n$. After establishing this result, one rushes to bring all the tools of matrix theory into vector spaces.

This is of course all correct, we offer now only a different emphasis. If you have a vector space V over the field $\mathbb{R}$ of dimension n then unlike $\mathbb{R}^n$ is does not have a distinguished ordered basis. Let us put our focus on the isomorphisms to $\mathbb{R}^n$ rather than the ordered bases. If you have two linear isomorphisms $\phi_1, \phi_2 : V \to \mathbb{R}^n$ then you can compose them to a linear isomorphism between euclidean spaces $\phi_2 \circ \phi_1^{-1} : \mathbb{R}^n \to \mathbb{R}^n$. It is the change of basis matrix that you learn about in vector space theory. We, being differential geometers, recognise this situation as a manifold V with coordinates ϕ_1, ϕ_2 and a transition function $\phi_2 \circ \phi_1^{-1}$. Just as for a C^k-manifold the transition maps are bijective C^k functions with C^k inverse, here the transition maps are bijective linear maps (which makes the inverse automatically linear). In this way, we could define a vector space as a set V with a *linear atlas*, an atlas where all transition maps are linear isomorphisms, as opposed to defining a vector space through a list of axioms. Thus a vector space is a special type of manifold and the choice of a basis is the choice of a chart.

Let us say that matrices describe linear maps between $\mathbb{R}^n$ and $\mathbb{R}^m$, but not other vector spaces. Suppose you have two vector spaces V, W of dimension n, m with linear isomorphisms $\phi_1, \phi_2 : V \to \mathbb{R}^n$ and $\psi_1, \psi_2 : W \to \mathbb{R}^m$ and a linear map $A : V \to W$. For any manifold, we can examine a map between manifolds in charts $\psi_1 \circ A \circ \phi_1^{-1} : \mathbb{R}^n \to V \to W \to \mathbb{R}^m$. This function $\psi_1 \circ A \circ \phi_1^{-1}$ is a linear function between euclidean spaces, unlike A, so it can be represented as a matrix. Thus writing a linear map as matrix with respect to bases is nothing other than writing it in charts. Likewise $\psi_2 \circ A \circ \phi_2^{-1}$ can also be represented as a matrix, and the relationship between the two is exactly that change of coordinates formula from differential geometry:

$$\psi_2 \circ A \circ \phi_2^{-1} = (\psi_2 \circ \psi_1^{-1}) \circ (\psi_1 \circ A \circ \phi_1^{-1}) \circ (\phi_2 \circ \phi_1^{-1})^{-1}.$$

We can interpret this formula in terms of matrices:

$$\begin{aligned}&\text{matrix of } A \text{ with respect to the new bases}\\ &\quad = \text{change of basis matrix for } W\\ &\quad \times \text{matrix of } A \text{ with respect to the old bases}\\ &\quad \times \text{inverse of change of basis matrix for } V.\end{aligned}$$

Hopefully you agree that by using the differential geometry idea of clearly separating an object from its coordinates/charts, we clarify how to change coordinates.

Let us consider Gaussian elimination, aka row and column operations. Perhaps you are aware that the three elementary row operations can be implemented as matrix multiplication. Here they are for $2 \times n$ matrix:

$$\begin{aligned}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}v_1 & \dots\\ w_1 & \dots\end{pmatrix} &= \begin{pmatrix}w_1 & \dots\\ v_1 & \dots\end{pmatrix}\\ \begin{pmatrix}1 & 0\\ 0 & \lambda\end{pmatrix}\begin{pmatrix}v_1 & \dots\\ w_1 & \dots\end{pmatrix} &= \begin{pmatrix}v_1 & \dots\\ \lambda w_1 & \dots\end{pmatrix}\\ \begin{pmatrix}1 & 0\\ \lambda & 1\end{pmatrix}\begin{pmatrix}v_1 & \dots\\ w_1 & \dots\end{pmatrix} &= \begin{pmatrix}v_1 & \dots\\ w_1 + \lambda v_1 & \dots\end{pmatrix}.\end{aligned}$$

We see that all three 2×2 matrices are linear isomorphisms, therefore can be interpreted as a change of basis (or a change of chart). In a similar way, the elementary column operations are the multiplication of an invertible matrix from the right. If we have a linear map $A : V \to W$ we see that starting with the matrix of A in some bases applying row and column operations gives CAD^{-1}, the matrix of A with respect to some other bases. The point of Gaussian elimination is to bring a matrix to reduced row echelon form. But this is exactly finding special bases (or charts) for V and W in which the matrix of A has a simple form, namely an identity block and the rest zeroes.

Consider if we have a linear map $A : V \to V$. We can treat the domain and codomain as separate and use Gaussian elimination to find two bases of V that together give the matrix of A a nice form. But we can also ask what can be done if we are forced to use the same basis of V for both domain and codomain. If the change of basis matrix is C, we are asking what can be said about CAC^{-1}. This is matrix conjugation/similarity. One learns that the eigenvalues of a matrix are preserved by conjugation and the complete answer is given by the Jordan normal form.

Bilinear Forms

Introductory linear algebra concentrates on vector spaces and linear maps. But these are not the only topics of linear algebra. In particular, bilinear forms were perhaps just as much studied as linear maps. A bilinear form is a function $B : V \times V \to \mathbb{R}$ that is linear in both arguments. The term 'form' is an old term indicating a function to the scalars. Like linear maps, bilinear forms on $\mathbb{R}^n$ can be represented as a square matrix $B(v, w) = v^T Bw$. What is different to linear maps is that under a change of basis C^{-1} the matrix of a bilinear form changes to $C^T BC$. This is

called matrix *congruence* and leads to a different set of invariants. For example, the eigenvalues are invariants of a linear map because they are roots of $\det(\lambda I - A)$ and

$$\det(\lambda I - CAC^{-1}) = \det C(\lambda I - A)C^{-1} = \det C^{-1}C(\lambda I - A) = \det(\lambda I - A).$$

The same calculation does not work if the transformation is $C^T BC$. Therefore eigenvalues are not invariants of matrices of bilinear forms.

Every bilinear form can be split into a symmetric and antisymmetric part

$$B(v,w) = \frac{1}{2}\Big(B(v,w) + B(w,v)\Big) + \frac{1}{2}\Big(B(v,w) - B(w,v)\Big).$$

A bilinear form is symmetric/antisymmetric if and only if the matrix is symmetric/antisymmetric. This is an invariant of the matrices of bilinear forms; a linear map might have a symmetric matrix in one basis but not another. Another invariant is the dimension of the left and right kernels of bilinear form. The left kernel is the set of vectors v such that $B(v,w) = 0$ for all $w \in V$, and ditto for the right kernel. These are vector subspaces and have the same dimension, called the *nullity*. For symmetric and antisymmetric bilinear forms, the left and right kernels are equal. A bilinear form is called *non-degenerate* if its kernels are trivial.

The theory for the symmetric and antisymmetric are significantly different from each other. The invariants of a symmetric bilinear form are revealed by Sylvester's law of inertia.

Theorem B.1 (Sylvester's law of inertia)**.** *There is a basis that reduces the matrix of a symmetric bilinear form B to the following block-diagonal matrix*

$$\begin{pmatrix} I_p & & \\ & -I_q & \\ & & 0 \end{pmatrix}.$$

The sizes (p,q) are called the signature and are invariants.

Proof. We first show that B can be diagonalised with $\{1,-1,0\}$ on the diagonal. If all vectors have $B(v,v) = 0$, then

$$0 = B(v+w, v+w) - B(v-w, v-w) = 4B(v,w)$$

shows that B is identically zero. Hence it is the zero matrix in any basis. Otherwise there is a vector with $B(v,v) \neq 0$. By rescaling

$$B\left(\sqrt{|B(v,v)|}^{-1} v, \sqrt{|B(v,v)|}^{-1} v\right) = \sqrt{|B(v,v)|}^{-2} B(v,v) = |B(v,v)|^{-1} B(v,v) = \pm 1.$$

So without loss of generality assume $B(v,v) = \pm 1$. Then we can consider its orthogonal complement $v^\perp := \{w \in V \mid B(v,w) = 0\}$. This is the kernel of $w \mapsto B(v,w)$. Its image is 1 dimensional by the definition of v, so $v^\perp$ is $(n-1)$-dimensional. We can restrict B to $v^\perp$ and apply the inductive hypothesis to get a basis $(v_2, \ldots, v_n)$ for which $B(v_i, v_j) = 0$ for $i \neq j$ and $B(v_i, v_i) \in \{1,-1,0\}$. But additionally $B(v, v_i) = 0$ and $B(v,v) \in \{1,-1\}$. So $(v, v_2, \ldots, v_n)$ is a basis that meets our requirements.

Clearly by reordering the basis we can ensure that all the 1s come first, then the -1s and lastly the 0s. It remains to show that signature is an invariant. The kernel of B is independent of

basis, so we know that the sum $p+q$ is an invariant r. Take any two diagonalising bases, v_i with signature (p,q) and w_i with (p',q'). Suppose that $p<p'$. Construct a linear map

$$L:\mathbb{R}^n\to\mathbb{R}^{p+q'},\quad x\mapsto\Big(B(v_1,x),\dots,B(v_p,x),B(w_{p'+1},x),\dots,B(w_{p'+q'},x)\Big).$$

Any vector $x\in\ker B$ is also in the kernel of L. Notice however that $p+q'<p'+q'=n-r$ so $\dim\ker L=n-(p+q')>r$. Hence there must be a vector u in the kernel of L that isn't in the kernel of B.

We write $u=\sum_i u_iv_i$. The coefficients with $i\le p$ must be zero, since for $j\le p$

$$0=B(v_j,u)=\sum_i u_iB(v_j,v_i)=u_j.$$

And because u is not in the kernel of B, at least one of the coefficients with $p+1\le i\le p+q$ must be non-zero. Therefore

$$B(u,u)=\sum_{i,j\ge p+1}u_iu_jB(v_i,v_j)=\sum_{i\ge p+1}u_iu_iB(v_i,v_i)=\sum_{p+1\le i\le p+q}u_i^2(-1)<0.$$

But now we can apply the same argument in the w_i basis, but this time with the conclusion that $B(u,u)>0$. This is a contradiction, proving that $p=p'$. □

A vector space with a non-degenerate symmetric bilinear form is called an inner product space, which should already be familiar to you. Non-degeneracy is equivalent to a signature of $(n,0)$. Hence all inner product spaces of the same finite dimension are essentially the same. The above proof in this case yields an orthonormal basis. In fact, we can insert the Gram-Schmidt process to make the above proof fully constructive. After we have normalised v to have $B(v,v)=\pm1$ choose more vectors to have a basis $(v,v_2',\dots v_n')$ of the whole vector space. Observe that

$$B\left(v,v_i'-\frac{B(v,v_i')}{B(v,v)}v\right)=B(v,v_i')-\frac{B(v,v_i')}{B(v,v)}B(v,v)=0.$$

Hence $v_i:=v_i'-\frac{B(v,v_i')}{B(v,v)}v$ is a basis of $v^\perp$.

Slightly more can be said about a symmetric bilinear form on an inner product space. This is a situation we often find ourselves in in Riemannian geometry, were the metric g is the inner product and we have another symmetric bilinear form such as the second fundamental form h. In particular, in this context we can identity a preferred class of bases, namely the orthonormal bases. The transformation from one orthonormal basis to another is an orthogonal matrix, so

$$\det(\lambda I-O^TAO)=\det O^T(\lambda I-A)O=\det OO^T(\lambda I-A)=\det(\lambda I-A).$$

The characteristic polynomial (and hence eigenvalues, determinant, trace) of the matrix of a symmetric bilinear form in an orthonormal basis is an invariant. This explains Lemma 1.34. We essentially defined the principal curvatures to be the eigenvalues of the second fundamental form (with respect to g-orthonormal bases). The Gauss curvature is their product, ie the determinant. The determinant can then be taken with respect to any orthonormal basis and be the same.

Finally, a vector space with a non-degenerate antisymmetric bilinear form is called an symplectic space. Antisymmetric bilinear forms can only be non-degenerate if the dimension is even. For

a symplectic space it is possible to find a basis (Darboux basis) such that the matrix of the bilinear form is

$$\begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}.$$

Just like for inner product spaces, there is essentially only one symplectic vector space in every even dimension. Symplectic vector spaces are the starting point for symplectic geometry in the same way that inner product spaces are the starting point for Riemannian geometry. Unlike in Riemannian geometry, which has curvature, symplectic geometry does not have local invariants.