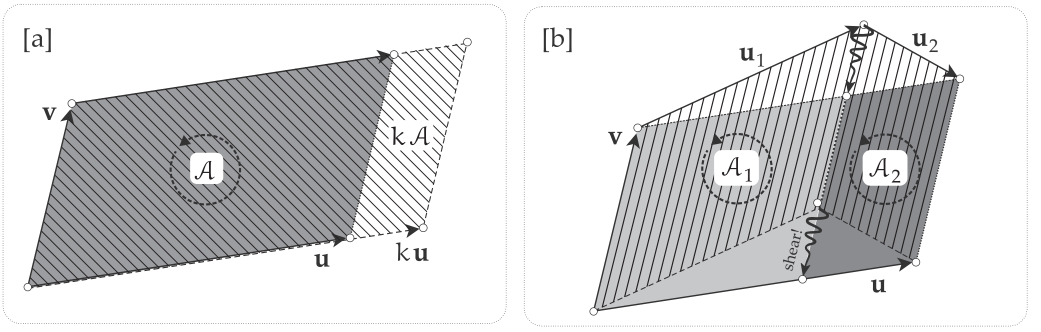
Contents



[34.1] Geometrical Proof That Oriented Area Is a 2-Form. **[a]** *Expanding an edge by* k *expands the area by* k*:* $\boldsymbol{\mathcal{A}}(\mathrm{k}\mathbf{u},\mathbf{v}) = \mathrm{k}\,\boldsymbol{\mathcal{A}}(\mathbf{u},\mathbf{v})$. **[b]** *Shearing* $\mathcal{A}_1 = \boldsymbol{\mathcal{A}}(\mathbf{u}_1,\mathbf{v})$ *and* $\mathcal{A}_2 = \boldsymbol{\mathcal{A}}(\mathbf{u}_2,\mathbf{v})$ *in the direction of* $\mathbf{v}$ *preserves their areas, so* $\mathcal{A} = \boldsymbol{\mathcal{A}}(\mathbf{u},\mathbf{v}) = \mathcal{A}_1 + \mathcal{A}_2$.

Only gradually will it become apparent the extent to which Élie Cartan was crazy like a fox. As was the case with 1-forms, so too with 2-forms: Cartan recognized that they had been hiding in plain sight!

34.2 Example: The Area 2-Form

In $\mathbb{R}^2$, let us define

$$\boxed{\boldsymbol{\mathcal{A}}(\mathbf{u},\mathbf{v}) = \text{oriented area of the parallelogram with edges } \mathbf{u} \text{ and } \mathbf{v}.}$$

Then $\boldsymbol{\mathcal{A}}$ is a 2-form!

First, it is immediately clear that $\boldsymbol{\mathcal{A}}(\mathbf{u},\mathbf{v})$ is antisymmetric: if we swap $\mathbf{u}$ and $\mathbf{v}$, then the magnitude of the area is unaltered, but the orientation of the parallelogram is reversed. It remains to verify that $\boldsymbol{\mathcal{A}}$ is a tensor, i.e., that it is *linear* in each slot. As we did with 1-forms, we may break down the linearity requirement into two parts: (32.2) and (32.3), applied to each slot.

The truth of (32.3) is explained by [34.1a], which illustrates the fact that if we expand either edge of a parallelogram by k, then its area is expanded by k, too:

$$\boldsymbol{\mathcal{A}}(\mathrm{k}\mathbf{u},\mathbf{v}) = \boldsymbol{\mathcal{A}}(\mathbf{u},\mathrm{k}\mathbf{v}) = \mathrm{k}\,\boldsymbol{\mathcal{A}}(\mathbf{u},\mathbf{v}).$$

Note that if $\boldsymbol{\Psi}$ is an arbitrary 2-form, then if (32.3) is true for one slot, it *must* be true for the other slot, too:

$$\boldsymbol{\Psi}(\mathrm{k}\mathbf{u},\mathbf{v}) = \mathrm{k}\,\boldsymbol{\Psi}(\mathbf{u},\mathbf{v}) \qquad \Longrightarrow \qquad \boldsymbol{\Psi}(\mathbf{u},\mathrm{k}\mathbf{v}) = -\boldsymbol{\Psi}(\mathrm{k}\mathbf{v},\mathbf{u}) = -\mathrm{k}\,\boldsymbol{\Psi}(\mathbf{v},\mathbf{u}) = \mathrm{k}\,\boldsymbol{\Psi}(\mathbf{u},\mathbf{v}).$$

The truth of (32.2) is less obvious, but the argument just given means that we need only prove it for the first slot—the truth for the second slot then follows from antisymmetry.

Let $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$, and define

$$\mathcal{A} = \boldsymbol{\mathcal{A}}(\mathbf{u},\mathbf{v}), \qquad \mathcal{A}_1 = \boldsymbol{\mathcal{A}}(\mathbf{u}_1,\mathbf{v}), \qquad \mathcal{A}_2 = \boldsymbol{\mathcal{A}}(\mathbf{u}_2,\mathbf{v}).$$

Then (32.2) requires that

$$\boxed{\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2.}$$

That this is indeed true is demonstrated geometrically in [34.1b]. Thus $\boldsymbol{\mathcal{A}}$ is indeed a 2-form.

ACT III
Curvature

ACT IV
Parallel Transport

ACT V
Forms

Chapter 1

Euclidean and Non-Euclidean Geometry

1.1 Euclidean and Hyperbolic Geometry

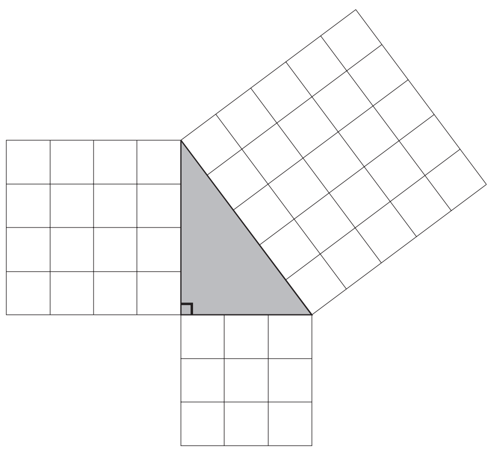
Differential Geometry is the application of calculus to the geometry of space that is *curved*. But to understand space that is curved we shall first try to understand space that is *flat*.

We inhabit a natural world pervaded by curved objects, and if a child asks us the meaning of the word "flat," we are most likely to answer in terms of the *absence* of curvature: a smooth surface *without* any bumps or hollows. Nevertheless, the very earliest mathematicians seem to have been drawn to the singular simplicity and uniformity of the flat plane, and they were rewarded with the discovery of startlingly beautiful facts about geometric figures constructed within it. With the benefit of enormous hindsight, some of these facts can be seen to *characterize* the plane's flatness.

One of the earliest and most profound such facts to be discovered was Pythagoras's Theorem. Surely the ancients must have been awed, as any sensitive person must remain today, that a seemingly unalloyed fact about *numbers*,

$$3^2+4^2=5^2,$$

in fact has *geometrical* meaning, as seen in [1.1].[1]



[1.1] *Pythagoras's Theorem: the geometry of* $3^2+4^2=5^2$.

While Pythagoras himself lived in Greece around 500 BCE, the theorem bearing his name was discovered much earlier, in various places around the world. The earliest known example of such knowledge is recorded in the Babylonian clay tablet (catalogued as "Plimpton 322") shown in [1.2], which was unearthed in what is now Iraq, and which dates from about 1800 BCE.

The tablet lists ***Pythagorean triples***:[2] integers (a, b, h) such that h is the hypotenuse of a right triangle with sides a and b, and therefore $a^2+b^2=h^2$. Some of these ancient examples are impressively large, and it seems clear that they did not stumble upon them, but rather possessed a mathematical process for generating solutions. For example, the fourth row of the tablet records the fact that $13500^2+12709^2=18541^2$.

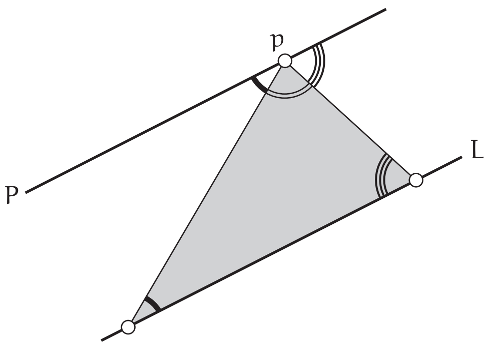
The deeper knowledge that underlay these ancient results is not known,[3] but to find the first evidence of the "modern," logical, deductive approach to mathematics we must jump 1200 years into the future of the clay tablet. Scholars believe that it was Thales of Miletus (around 600 BCE)

[1] We repeat what was said in the Prologue: *equations* are labelled with parentheses (*round* brackets)—(...), while *figures* are labelled with *square* brackets—[...].

[2] In fact the tablet only records *two* members (a, h) of the triples (a, b, h).

[3] In the seventeenth century, Fermat and Newton reconstructed and generalized a *geometrical* method of generating the general solution, due to Diophantus. See Exercise 5.

[1.2] *Plimpton 322: A clay tablet of Pythagorean triples from 1800 BCE.*



[1.3] *Euclid's Parallel Axiom:* P *is the* unique *parallel to* L *through* p*, and the angle sum of a triangle is* π.

who first pioneered the idea of deducing new results from previously established ones, the logical chain beginning at a handful of clearly articulated assumptions, or ***axioms***.

Leaping forward again, 300 years beyond Thales, we find one of the most perfect exemplars of this new approach in Euclid's *Elements*, dating from 300 BCE. This work sought to bring order, clarity, and rigour to geometry by deducing everything from just five simple axioms, the fifth and last of which dealt with parallel lines.

Defining two lines to be ***parallel*** if they do not meet, Euclid's Fifth Axiom[4] is illustrated in [1.3]:

> **Parallel Axiom.** *Through any point* p *not on the line* L *there exists precisely one line* P *that is parallel to* L.

But the character of this axiom was more complex and less immediate than that of the first four, and mathematicians began a long struggle to dispense with it as an assumption, instead seeking to show that it must be a logical *consequence* of the first four axioms.

This tension went unresolved for the next *2000 years*. As the centuries passed, many attempts were made to prove the Parallel Axiom, and the number and intensity of these efforts reached a crescendo in the 1700s, but all met with failure.

Yet along the way useful *equivalents* of the axiom emerged. For example: *There exist similar triangles of different sizes* (Wallis in 1663; see Stillwell (2010)). But the very first equivalent was already present in Euclid, and it is the one still taught to every school child: *the angles in a triangle add up to two right angles.* See [1.3].

The explanation of these failures only emerged around 1830. Completing a journey that had begun 4000 years earlier, Nikolai Lobachevsky and János Bolyai independently announced the

[4]Euclid did not state his axiom in this form, but it is logically equivalent.

discovery of an entirely new form of geometry (now called ***Hyperbolic Geometry***) taking place in a new kind of plane (now called the ***hyperbolic plane***). In this Geometry the first four Euclidean axioms hold, but the parallel axiom does *not*. Instead, the following is true:

> ***Hyperbolic Axiom.*** *There are at least two parallel lines through* p *that do not meet* L. (1.1)

These pioneers explored the logical consequences of this axiom, and by purely abstract reasoning were led to a host of fascinating results within a rich new geometry that was bizarrely different from that of Euclid.

Many others before them, perhaps most notably Saccheri (in 1733; see Stillwell 2010) and Lambert (in 1766; see Stillwell 2010), had discovered some of these consequences of (1.1), but their aim in exploring these consequences had been to find a *contradiction*, which they believed would finally prove that Euclidean Geometry to be the One True Geometry.

[1.4] *Johann Heinrich Lambert (1728–1777).*

Certainly Saccheri believed he had found a clear contradiction when he published “Euclid Freed of Every Flaw.” But Lambert is a much more perplexing case, and he is perhaps an unsung hero in this story. His results penetrated so deeply into this new geometry that it seems impossible that he did not at times believe in the reality of what he was doing. Regardless of his motivation and beliefs[5], Lambert (shown in [1.4]) was certainly the first to discover a remarkable fact[6] about the angle sum of a triangle under axiom (1.1), and his result will be central to much that follows in Act II.

Nevertheless, Lobachevsky and Bolyai richly deserve their fame for having been the first to recognize and fully embrace the idea that they had discovered an entirely new, consistent, non-Euclidean Geometry. But what this new geometry really *meant*, and what it might be useful for, even they could not say.[7]

Remarkably and surprisingly, it was the *Differential Geometry of curved surfaces* that ultimately resolved these questions. As we shall explain, in 1868 the Italian mathematician Eugenio Beltrami finally succeeded in giving Hyperbolic Geometry a concrete interpretation, setting it upon a firm and intuitive foundation from which it has since grown and flourished. Sadly, neither Lobachevsky nor Bolyai lived to see this: they died in 1856 and 1860, respectively.

This non-Euclidean Geometry had in fact already manifested itself in various branches of mathematics throughout history, but always in disguise. Henri Poincaré (beginning around 1882) was the first not only to strip it of its camouflage, but also to recognize and exploit its power

---

[5]I thank Roger Penrose for making me see that Lambert deserves greater credit than he is usually granted. Penrose did so by means of the following analogy: “Should we not give credit to Einstein for the cosmological constant, even if he introduced it for the wrong reasons? And should we blame him for later retracting it, calling it the “greatest blunder of my life”? Or what about General Relativity itself, which Einstein seemed to become less and less convinced was the right theory (needing to be replaced by some kind of non-singular unified field theory) as time went on?” [Private communication.]

[6]If you cannot wait, it’s (1.8).

[7]Lobachevsky did in fact put this geometry to use to evaluate previously unknown integrals, but (at least in hindsight) this particular application must be viewed as relatively minor.

in such diverse areas as Complex Analysis, Differential Equations, Number Theory, and Topology. Its continued vitality and centrality in the mathematics of the 20th and twenty-first centuries is demonstrated by Thurston's work on 3-manifolds, Wiles's proof of ***Fermat's Last Theorem***, and Perelman's proof of the ***Poincaré Conjecture*** (as a special case of Thurston's ***Geometrization Conjecture***), to name but three examples.

In Act II we shall describe Beltrami's breakthrough, as well as the nature of Hyperbolic Geometry, but for now we wish to explore a different, simpler kind of non-Euclidean Geometry, one that was already known to the Ancients.

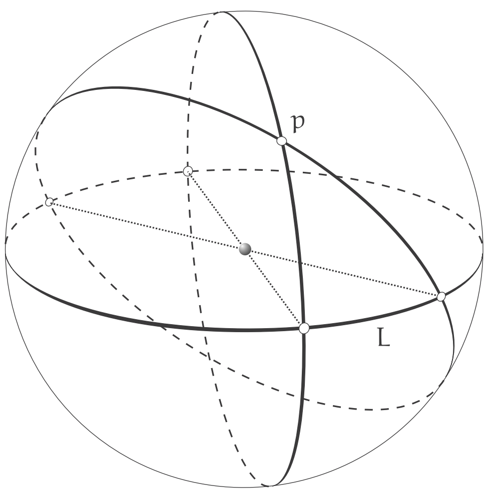
1.2 Spherical Geometry

To construct a non-Euclidean Geometry we must deny the existence of a unique parallel. The Hyperbolic Axiom assumes two or more parallels, but there is one other logical possibility—*no* parallels:

> ***Spherical Axiom.*** *There are* no *lines through* p *that are parallel to* L *:* every *line meets* L. (1.2)

Thus there are actually *two* non-Euclidean[8] geometries: spherical and hyperbolic.

As the name suggests, Spherical Geometry can be realized on the surface of a sphere—denoted $\mathbb{S}^2$ in the case of the *unit* sphere—which we may picture as the surface of the Earth. On this sphere, what should be the analogue of a "straight line" connecting two points on the surface? Answer: the shortest route between them! But if you wish to sail or fly from London to New York, for example, what *is* the shortest route?



[1.5] *The great circles of* $\mathbb{S}^2$ *intersect in pairs of antipodal points.*

The answer, already known to the ancient mariners, is that the shortest route is an arc of a ***great circle***, such as the equator, obtained by cutting the sphere with a plane passing through its centre. In [1.5] we have chosen L to be the equator, and it is clear that (1.2) is satisfied: every line through p meets L in a pair of ***antipodal*** (i.e., diametrically opposite) points.

In the plane, the shortest route is also the *straightest* route, and in fact the same is true on the sphere: in a precise sense to be discussed later, the great circle trajectory bends neither to the right nor to the left as it traverses the spherical surface.

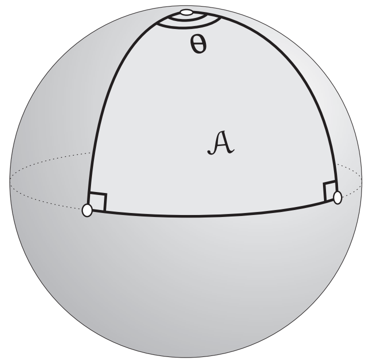
There are other ways of constructing the great circles on the Earth that do not require thinking about planes passing through the completely inaccessible centre of the Earth. For example, on a globe you may map out your great circle journey by holding down one end of a piece of string on London and pulling the string tightly over the surface so that the other end is on New York. The taut string has

[8] Nevertheless, the reader should be aware that in modern usage "non-Euclidean Geometry" is usually synonymous with "Hyperbolic Geometry."

automatically found the shortest, straightest route—the shorter[9] of the two arcs into which the great circle through the two cities is divided by those cities.

With the analogue of straight lines now found, we can "do geometry" within this spherical surface. For example, given three points on the surface of the Earth, we can connect them together with arcs of great circles to obtain a "triangle." Figure [1.6] illustrates this in the case where one vertex is located at the north pole, and the other two are on the equator.

But if this non-Euclidean Spherical Geometry was already used by ancient mariners to navigate the oceans, and by astronomers to map the spherical night sky, what then was so shocking and new about the non-Euclidean geometry of Lobachevsky and Bolyai?



[1.6] *A particularly simple "triangle" on the sphere.*

The answer is that this Spherical Geometry was merely considered to be inherited from the *Euclidean* Geometry of the 3-dimensional space in which the sphere resides. No thought was given in those times to the sphere's internal 2-dimensional geometry as representing an alternative to Euclid's plane. Not only did it violate Euclid's fifth axiom, it also violated a much more basic one (Euclid's first axiom) that we can always draw a unique straight line connecting two points, for this fails when the points are antipodal.

On the other hand, the Hyperbolic Geometry of Lobachevsky and Bolyai was a much more serious affront to Euclidean Geometry, containing familiar lines of infinite length, yet flaunting multiple parallels, ludicrous angle sums, and many other seemingly nonsensical results. Yet the 21-year-old Bolyai was confident and exuberant in his discoveries, writing to his father, *"From nothing I have created another entirely new world."*

We end with a tale of tragedy. Bolyai's father was a friend of Gauss, and sent him what János had achieved. By this time Gauss had himself made some important discoveries in this area, but had kept them secret. In any case, János had seen further than Gauss. A kind word in public from Gauss, the most famous mathematician in the world, would have assured the young mathematician a bright future. But Nature and nurture sometimes conspire to pour extraordinary mathematical gifts into a vessel marred by very ordinary human flaws, and Gauss's reaction to Bolyai's marvellous discoveries was mean-spirited and self-serving in the extreme.

First, Gauss kept Bolyai in suspense for six months, then he replied as follows:

> Now something about the work of your son. You will probably be shocked for a moment when I begin by saying that *I cannot praise it*, but I cannot do anything else, since to praise it would be to praise myself. The whole content of the paper, the path that your son has taken, and the results to which he has been led, agree almost everywhere with my own meditations, which have occupied me in part for 30–35 years.

Gauss did however "thank" Bolyai's son for having "saved him the trouble"[10] of having to write down theorems he had known for decades.

János Bolyai never recovered from the surgical blow delivered by Gauss, and he abandoned mathematics for the rest of his life.[11]

---

[9] If the two points are antipodal, such as the north and south poles, then the two arcs are the *same* length. Furthermore, the great circle itself is no longer unique: *every* meridian is a great circle connecting the poles.

[10] Gauss had previously denigrated Abel's discovery of elliptic functions in precisely the same manner; see Stillwell (2010, p. 236).

[11] If this depresses you, turn your thoughts to the uplifting counterweight of Leonhard Euler. An intellectual volcano erupting with wildly original thoughts (some of which we shall meet later) he was also a kind and generous spirit. We cite one, parallel

1.3 The Angular Excess of a Spherical Triangle

As we have said, the parallel axiom is equivalent to the fact that the angles in a triangle sum to π. It follows that both the spherical axiom and the hyperbolic axiom must lead to geometries in which the angles do *not* sum to π. To quantify this departure from Euclidean Geometry, we introduce the ***angular excess***, defined to be the amount $\mathcal{E}$ by which the angle sum exceeds π:

$$\boxed{\mathcal{E} \equiv (\text{angle sum of triangle}) - \pi.}$$

For example, for the triangle shown in [1.6], $\mathcal{E} = (\theta + \frac{\pi}{2} + \frac{\pi}{2}) - \pi = \theta$.

A crucial insight now arises if we compare the triangle's angular excess with its area $\mathcal{A}$. Let the radius of the sphere be R. Since the triangle occupies a fraction $(\theta/2\pi)$ of the northern hemisphere, $\mathcal{A} = (\theta/2\pi)\, 2\pi R^2 = \theta\, R^2$. Thus,

$$\boxed{\mathcal{E} = \frac{1}{R^2}\,\mathcal{A}.} \tag{1.3}$$

[1.7] *Thomas Harriot (1560–1621).*

In 1603 the English mathematician Thomas Harriot (see [1.7]) made the remarkable discovery[12] that this relationship holds for *any* triangle Δ on the sphere; see [1.8a]. Harriot's elementary but ingenious argument[13] goes as follows.

Prolonging the great-circle sides of Δ divides the surface of the sphere into eight triangles, the four triangles labelled Δ, Δ_α, Δ_β, Δ_γ each being paired with a congruent antipodal triangle. This is clearer in [1.8b]. Since the area of the sphere is $4\pi R^2$, we deduce that

$$\mathcal{A}(\Delta) + \mathcal{A}(\Delta_\alpha) + \mathcal{A}(\Delta_\beta) + \mathcal{A}(\Delta_\gamma) = 2\pi R^2. \tag{1.4}$$

On the other hand, it is clear in [1.8b] that Δ and Δ_α together form a wedge whose area is a fraction $(\alpha/2\pi)$ of the area of the sphere:

$$\mathcal{A}(\Delta) + \mathcal{A}(\Delta_\alpha) = 2\alpha R^2.$$

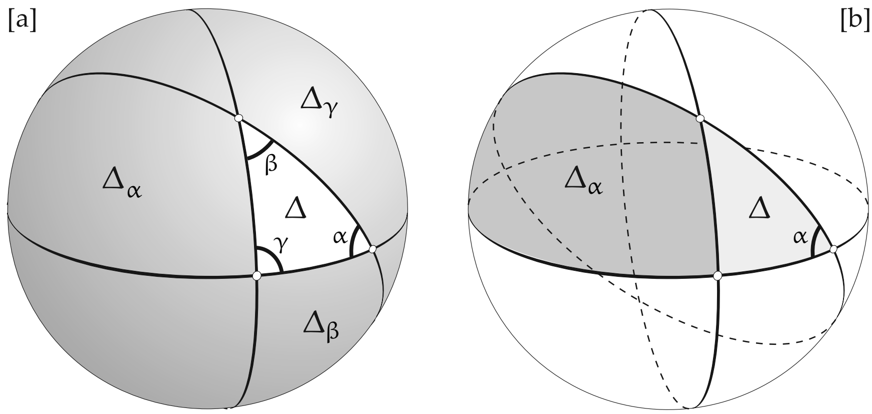
Similarly,

$$\begin{aligned} \mathcal{A}(\Delta) + \mathcal{A}(\Delta_\beta) &= 2\beta R^2, \\ \mathcal{A}(\Delta) + \mathcal{A}(\Delta_\gamma) &= 2\gamma R^2. \end{aligned}$$

example. When the then-unknown 19-year-old Lagrange wrote to him with overlapping discoveries in the calculus of variations, Euler wrote back: "... I deduced this myself. However, I decided to conceal this until you publish your results, since in no way do I want to take away from you any of the glory that you deserve." See Gindikin (2007, p. 216). Incidentally, Euler also personally intervened to rescue Lambert's career!

[12]This discovery is most often attributed to Girard, who rediscovered it about 25 years later.

[13]This argument was later rediscovered by Euler in 1781.



[1.8] *Harriot's Theorem (1603):* $\mathcal{E}(\Delta) = \mathcal{A}(\Delta)/R^2$.

Adding these last three equations, we find that

$$3\mathcal{A}(\Delta) + \mathcal{A}(\Delta_\alpha) + \mathcal{A}(\Delta_\beta) + \mathcal{A}(\Delta_\gamma) = 2(\alpha + \beta + \gamma)R^2. \tag{1.5}$$

Finally, subtracting (1.4) from (1.5), we find that

$$\mathcal{A}(\Delta) = R^2(\alpha + \beta + \gamma - \pi) = R^2\,\mathcal{E}(\Delta),$$

thereby proving (1.3).

1.4 Intrinsic and Extrinsic Geometry of Curved Surfaces

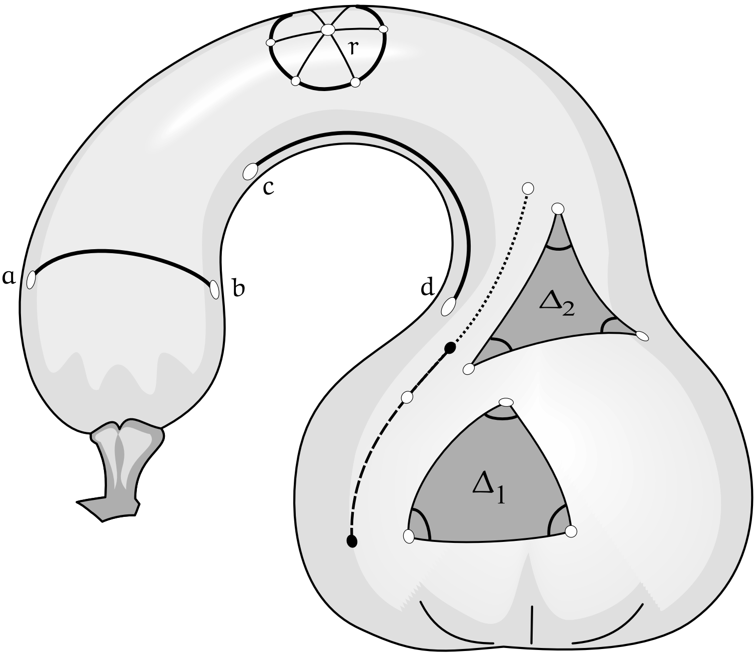
The mathematics associated with this stretched-string construction of a "straight line" will be explored in depth later in the book. For now we merely observe that the construction can be applied equally well to a *non*spherical surface, such as the crookneck squash shown in [1.9].

Just as on the sphere, we stretch a string over the surface, thereby finding the shortest, straightest route between two points, such as a and b. Provided that the string can slide around on the surface easily, the tension in the string ensures that the resulting path is as short as possible. Note that in the case of cd, we must imagine that the string runs over the *inside* of the surface.

In order to deal with all possible pairs of points in a uniform way, it is therefore best to imagine the surface as made up of two thinly separated layers, with the string trapped between them. On the other hand, this is only useful for thought experiments, not actual experiments. We shall overcome this obstacle shortly by providing a *practical* method of constructing these straightest curves on the surface of a physical object, even if the patch of surface bends the wrong way for a string to be stretched tightly over the outside of the object.

These shortest paths on a curved surface are the equivalent of straight lines in the plane, and they will play a crucial role throughout this book—they are called ***geodesics***. Thus, to use this new word, we may say that geodesics in the plane are straight lines, and geodesics on the sphere are great circles.

But even on the sphere the *length-minimizing* definition of "geodesic" is provisional, because we see that nonantipodal points are connected by *two* arcs of the great circle passing through them: the short one (which *is* the shortest route) and the long one. Yet the long arc is every bit as much a geodesic as the short one. There is the additional complication on the sphere that antipodal points



[1.9] *The* **intrinsic geometry** *of the surface of a crookneck squash:* **geodesics** *are the equivalents of straight lines, and triangles formed out of them may possess an angular excess of either sign, depending on how the surface bends:* $\mathcal{E}(\Delta_1) > 0$ *and* $\mathcal{E}(\Delta_2) < 0$.

are connected by *multiple* geodesics, and this *nonuniqueness* occurs on more general surfaces, too. What *is* true is that any two points that are *sufficiently close together* can be joined by a unique geodesic segment that is the shortest route between them.

Just as a line segment in the plane can be extended indefinitely in either direction by laying down overlapping segments, so too can a geodesic segment be extended on a curved surface, and this extension is unique. For example, in [1.9] we have extended the dashed geodesic segment connecting the black dots, by laying down the overlapping dotted segment between the white points.

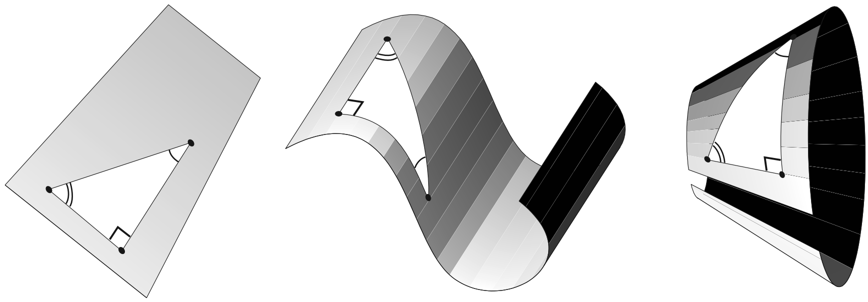
Because of the subtleties associated with the length-minimizing characterization of geodesics, before long we will provide an alternative, purely *local* characterization of geodesics, based on their straightness.

With these caveats in place, it is now clear how we should define distance within a surface such as [1.9]: the distance between two sufficiently close points a and b is the length of the geodesic segment connecting them.

Figure [1.9] shows how we may then define, for example, a "circle of radius r and centre c" as the locus of points at distance r from c. To construct this ***geodesic circle*** we may take a piece of string of length r, hold one end fixed at c, then (keeping the string taut) drag the other end round on the surface. But just as the angles in a triangle no longer sum to π, so now the circumference of a circle no longer is equal to $2\pi r$. In fact you should be able to convince yourself that for the illustrated circle the circumference is *less* than $2\pi r$.

Given three points on the surface, we may join them with geodesics to form a ***geodesic triangle***; [1.9] shows two such triangles, Δ_1 and Δ_2:

- Looking at the angles in Δ_1, it seems clear that they sum to *more* than π, so $\mathcal{E}(\Delta_1) > 0$, like a triangle in Spherical Geometry.



[1.10] *Bending a piece of paper changes the extrinsic geometry, but not the intrinsic geometry.*

- On the other hand, it is equally clear that the angles of Δ_2 sum to *less* than π: $\mathcal{E}(\Delta_2) < 0$, and (as we shall explain) this opposite behaviour is in fact exhibited by triangles in *Hyperbolic Geometry*. Note also that if we construct a circle in this saddle-shaped part of the surface, the circumference is now *greater* than $2\pi r$.

The concept of a geodesic belongs to the so-called ***intrinsic geometry*** of the surface—a fundamentally new view of geometry, introduced by Gauss (1827). It means the geometry that is knowable to tiny, ant-like, intelligent (but 2-dimensional!) creatures living *within* the surface. As we have discussed, these creatures can, for example, define a geodesic "straight line" connecting two nearby points as the shortest route within their world (the surface) connecting the two points. From there they can go on to define triangles, and so on. Defined in this way, it is clear that the intrinsic geometry is unaltered when the surface is bent (as a piece of paper can be) into quite different shapes in space, as long as distances *within* the surface are not stretched or distorted in any way. To the ant-like creatures within the surface, such changes are utterly undetectable.

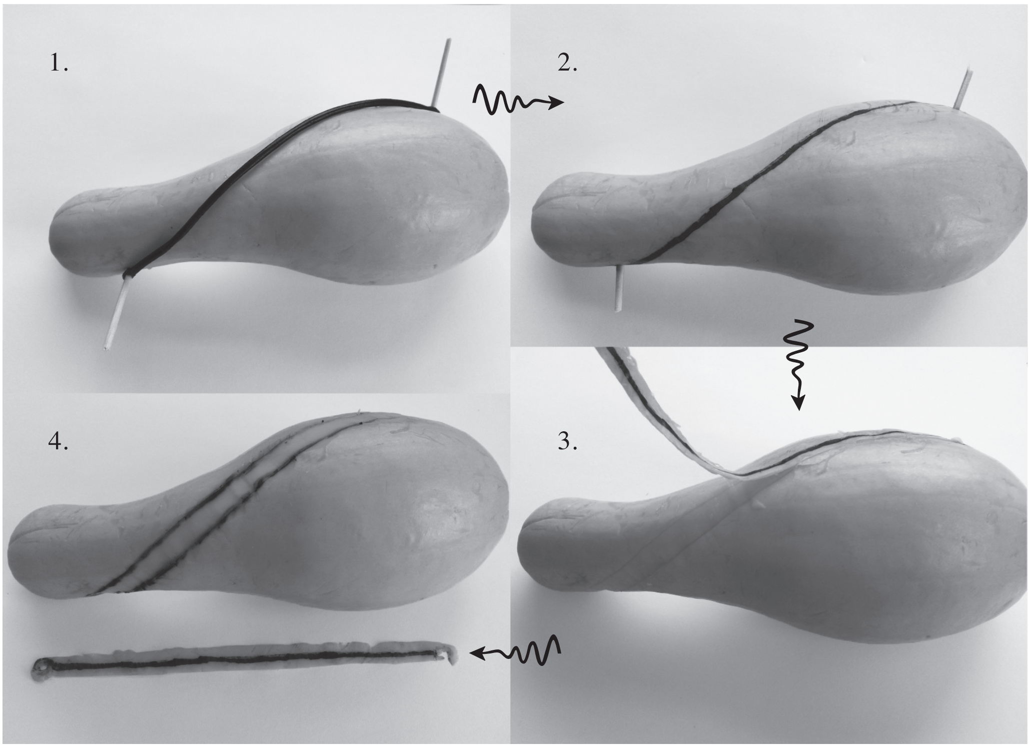
Under such a bending, the so-called ***extrinsic geometry*** (how the surface sits in space) most certainly does change. See [1.10]. On the left is a flat piece of paper on which we have drawn a triangle Δ with angles $(\pi/2)$, $(\pi/6)$, and $(\pi/3)$. Of course $\mathcal{E}(\Delta) = 0$. Clearly we can bend such a flat piece of paper into either of the two (extrinsically) curved surfaces on the right.[14] However, *intrinsically* these surfaces have undergone no change at all—they are both as flat as a pancake! The illustrated triangles on these surfaces (into which Δ is carried by our stretch-free bending of the paper) are identical to the ones that intelligent ants would construct using geodesics, and in both cases $\mathcal{E} = 0$: geometry on these surfaces is Euclidean.

Even if we take a patch of a surface that is intrinsically *curved*, so that a triangle within it has $\mathcal{E} \neq 0$, it too can generally be bent somewhat without stretching or tearing it, thereby altering its extrinsic geometry while leaving its intrinsic geometry unaltered. For example, cut a ping pong ball in half and gently squeeze the rim of one of the hemispheres, distorting that circular rim into an oval (but not an oval lying in a single plane).

1.5 Constructing Geodesics via Their Straightness

We have already alluded to the fact that geodesics on a surface have at least *two* characteristics in common with lines in the plane: (1) they provide the *shortest* route between two points that are not too far apart *and* (2) they provide the "straightest" route between these points. In this section we seek to clarify what we mean by "straightness," leading to a very simple and *practical* method of constructing geodesics on a physical surface.

[14] But note that we must first trim the edges of the rectangle to bend it into the shape on the far right.



[1.11] *On the curved surface of a fruit or vegetable, peel a narrow strip surrounding a geodesic, then lay it flat on the table. You will obtain a* straight line *in the plane!*

Most texts on Differential Geometry pay scant attention to such practical matters, and it is perhaps for this reason that the construction we shall describe is surprisingly little known in the literature.[15] In sharp contrast, in this book we *urge* you to explore the ideas by all means possible: theoretical contemplation, drawing, computer experiments, and (especially!) physical experiments with actual surfaces. Your local fruit and vegetable shop can supply your laboratory with many interesting shapes, such as the yellow summer squash shown in [1.11].

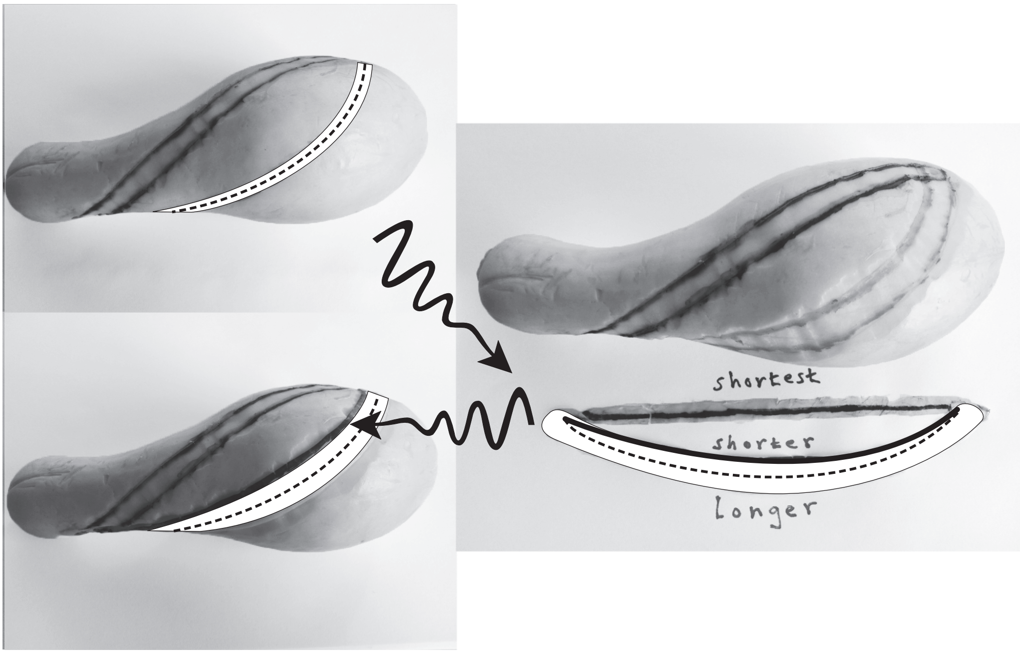
We can now use this vegetable to reveal the hidden straightness of geodesics via an experiment that we hope you will repeat for yourself:

1. On a fruit or vegetable, construct a geodesic by stretching a string over its curved surface.
2. Use a pen to trace the path of the string, then remove the string.
3. Make shallow incisions on either side of (and close to) the inked path, then use a vegetable peeler or small knife to remove the narrow strip of peel between the two cuts.
4. Lay the strip of peel flat on the table, and witness the marvellous fact that the geodesic within the peeled strip has become a *straight line* in the plane!

But why?!

To understand this, first let us be clear that although the strip is free to bend in the direction perpendicular to the surface (i.e., perpendicular to itself), it is *rigid* if we try to bend it sideways, tangent to the surface. Now let us employ proof by contradiction, and imagine what would happen if such a peeled geodesic did *not* yield a straight line when laid flat on the table. It is both a

[15]One of the rare exceptions is Henderson (1998), which we strongly recommend to you; for more details, see the *Further Reading* section at the end of this book.



[1.12] *Suppose that the illustrated dotted path is a geodesic such that a narrow (white) strip surrounding it does* not *become a straight line when laid flat in the plane. But in that case we can shrink the dotted path in the plane (towards the shortest, straight-line route in the plane) thereby producing the solid path. But if we then reattach the strip to the surface, this solid path is still shorter than the original dotted path, which was supposed to be the shortest path within the surface—a contradiction!*

drawback and an advantage of conducting such physical experiments that they will simply not *permit* us to construct something that is impossible, as is required in our desired mathematical proof by contradiction. Nevertheless, let us *suppose* that there exists a geodesic path, such as the dotted one shown on top left of [1.12], that when peeled and laid flat on the table (on the right) does *not* become a straight line.

The shortest route between the ends of this dotted (nonstraight) plane curve is the straight line connecting them. (As illustrated, this is the path of the *true* geodesic we already found using the string—but pretend you don't know that for now!) Thus we may shorten the dotted curve by deforming it slightly towards this straight, shortest route, yielding the solid path along the edge of the peeled strip. Therefore, after reattaching the strip to the surface (bottom left) the solid curve provides a shorter route over the surface than the dotted one, which we had supposed to be the shortest: a contradiction! Thus we have proved our previous assertion:

> *If a narrow strip surrounding a segment* G *of a geodesic is cut out of a surface and laid flat in the plane, then* G *becomes a segment of a* straight line. (1.6)

We are now very close to the promised simple and practical construction of geodesics. Look again at step 3 of [1.11], where we peeled off the strip of surface. But imagine now that we are *reattaching* the strip to the surface, instead. Ignore the history of how we got to this point: what are we actually doing right now in this reattachment process? We have picked up a narrow straight strip (of three-dimensional peel—but mathematically idealized as a two-dimensional strip) and we have unrolled it back onto the surface into the shallow channel from which we cut it. But here is the crucial observation: this shallow channel need not exist—the *surface* decides where the strip must go as we unroll it!

Thus, as a kind of time-reversed converse of (1.6), we obtain a remarkably simple and practical method[16] of constructing geodesics on a physical surface:

To construct a geodesic on a surface, emanating from a point p *in direction* $\mathbf{v}$, *stick one end of a length of narrow sticky tape down at* p *and unroll it onto the surface, starting in the direction* $\mathbf{v}$. (1.7)

(Note, however, that this does *not* provide a construction of the geodesic connecting p to a specified *target* point q.)

If this construction seems too simple to be true, please try it on any curved surface you have to hand. You can check that the sticky tape[17] is indeed tracing out a geodesic by stretching a string over the surface between two points on the tape: the string will follow the same path as the tape. But note that, as a promised bonus, this new tape construction works on *any* part of a surface, even where the surface is concave towards you, so that the stretched-string construction breaks down.

Of course all of this is a concrete manifestation of a mathematical idealization. A totally flat narrow strip of tape of nonzero width *cannot*[18] be made to fit perfectly on a genuinely curved surface, but its centre line *can* be made to rest on the surface, while the rest of the tape is tangent to the surface.

1.6 The Nature of Space

Let us return to the history of the discovery of non-Euclidean Geometry, and take our first look at how these two new geometries differ from Euclid's.

As we have said, Euclidean Geometry, is characterized by the vanishing of $\mathcal{E}(\Delta)$. Note that, unlike the original formulation of the parallel axiom, *this statement can be checked against experiment*: construct a triangle, measure its angles, and see if they add up to π. Gauss may have been the first person to ever conceive of the possibility that physical space might not be Euclidean, and he even attempted the above experiment, using three mountain tops as the vertices of his triangle, and using light rays for its edges.

Within the accuracy permitted by his equipment, he found $\mathcal{E}=0$. Quite correctly, Gauss did not conclude that physical space is definitely Euclidean in structure, but rather that if it is *not* Euclidean then its deviation from Euclidean Geometry is extremely small. But he did go so far as to say (see Rosenfeld 1988, p. 215) that he wished that this non-Euclidean Geometry might apply to the real world. In Act IV we shall see that this was a prophetic statement.

[16] This important fact is surprisingly hard to find in the literature. After we (re)discovered it, more than 30 years ago, we began searching, and the earliest mention of the underlying idea we could find at that time was in Aleksandrov (1969, p. 99), albeit in a less practical form: he imagined pressing a flexible metal ruler down onto the surface. Later, the basic idea also appeared in Koenderink (1990), Casey (1996), and Henderson (1998). *However*, we have since learned that the essential idea (though not in our current, *practical* form) goes all the way back to Levi-Civita, more than a century ago! See the footnote on page 236.

[17] We recommend using masking tape (aka painter's tape) because it comes in bright colours, and once a strip has been created, it can be detached and reattached repeatedly, with ease. A simple way to create narrow strips (from the usually wide roll of tape) is to stick a length of tape down onto a kitchen cutting board, then use a sharp knife to cut down its length, creating strips as narrow as you please.

[18] This is a consequence of a fundamental theorem we shall meet later, called the *Theorema Egregium*.

Although Gauss had bragged to friends that he had anticipated the Hyperbolic Geometry of Lobachevsky and Bolyai by decades, even he had unknowingly been scooped on some of its central results.

In 1766 (eleven years before Gauss was born) Lambert rediscovered Harriot's result on the sphere and then broke totally new ground in pursuing the analogous consequences of the Hyperbolic Axiom (1.1). First, he found that a triangle in Hyperbolic Geometry (if such a thing even existed) would behave *oppositely* to one in Spherical Geometry:

- In Spherical Geometry the angle sum of a triangle is greater than π: $\mathcal{E} > 0$.
- In Hyperbolic Geometry the angle sum of a triangle is less than π: $\mathcal{E} < 0$.

Thus a hyperbolic triangle behaves like a triangle drawn on a saddle-shaped piece of surface, like Δ_2 in [1.9]. Later we shall see that this is no accident.

Furthermore, Lambert discovered the crucial fact that $\mathcal{E}(\Delta)$ is again simply proportional to $\mathcal{A}(\Delta)$:

In both Spherical and Hyperbolic Geometry,

$$\mathcal{E}(\Delta) = \mathcal{K}\,\mathcal{A}(\Delta), \tag{1.8}$$

where $\mathcal{K}$ is a constant that is positive in Spherical Geometry, and negative in Hyperbolic Geometry.

Several interesting observations can be made in connection with this result:

- Although there are no qualitative differences between them, there are nevertheless *infinitely many* different Spherical Geometries, depending on the value of the positive constant $\mathcal{K}$. Likewise, each negative value of $\mathcal{K}$ yields a different Hyperbolic Geometry.
- Since the angle sum of a triangle cannot be negative, $\mathcal{E} \geqslant -\pi$. Thus in Hyperbolic Geometry ($\mathcal{K} < 0$) we have the strange and surprising result that *no triangle can have an area greater than* $|\pi/\mathcal{K}|$.
- From (1.8) we deduce that two triangles of different size cannot have the same angles. In other words, in non-Euclidean Geometry, *similar triangles do not exist!* (This accords with Wallis's 1663 discovery that the existence of similar triangles is equivalent to the Parallel Axiom.)
- Closely related to the previous point is the fact that in non-Euclidean Geometry there exists an *absolute unit of length*. (Gauss himself found it to be an exciting possibility that this purely mathematical fact might be realized in the physical world.) For example, in Spherical Geometry we could define this absolute unit of length to be the side of *the* equilateral triangle having, for instance, angle sum 1.01π. Similarly, in Hyperbolic Geometry we could define it to be the side of *the* equilateral triangle having angle sum 0.99π.
- A somewhat more natural way of defining the absolute unit of length is in terms of the constant $\mathcal{K}$. Since the radian measure of angle is defined as a ratio of lengths, $\mathcal{E}$ is a pure number. On the other hand, the area $\mathcal{A}$ has units of (length)^2. It follows that $\mathcal{K}$ must have units of $1/$(length)^2, and so there exists a length R such that $\mathcal{K}$ can be written as follows: $\mathcal{K} = +(1/R^2)$ in Spherical Geometry; $\mathcal{K} = -(1/R^2)$ in Hyperbolic Geometry. Of course in Spherical Geometry we already know that the length R occurring in the formula $\mathcal{K} = +(1/R^2)$ is simply the

radius of the sphere. Later we will see that this length R occurring in the formula $\mathcal{K} = -(1/R^2)$ can be given an equally intuitive and *concrete* interpretation in Hyperbolic Geometry.

- The smaller the triangle, the harder it is to distinguish it from a Euclidean triangle: only when the linear dimensions are a significant fraction of R will the differences become discernable. For example, we humans are small compared to the radius of the Earth, so if we find ourselves in a boat in the middle of a lake, its surface appears to be a Euclidean plane, whereas in reality it is part of a sphere. This Euclidean illusion for small figures is the reason that Gauss chose the largest possible triangle to conduct his light-ray experiment, thereby increasing his chances of detecting any small curvature that might be present in the space through which the light rays travelled.

Index